The Need of Business Analytics in Today's Scenario

S. Yamini

Abstract: ***Analytics is having the huge potential to improve any business. Due to the increase in the technological developments, the data is collected, interpreted and analyzed everywhere. The power and the impact of the data is huge. Previously, the Managerial decisions were taken based on the experience or by the gut insights. But now, the Data Analytics is made it easy to improve the possibilities of making better decisions based on data. The primary focus of this research article is to summarize about how to work with data to make better managerial decisions. This article introduces about the statistical concepts in the right flow which will drive to inferences about the population regardless of the tools used.***

Keywords : Summary Statistics, Exploratory Data Analysis, Sampling, Hypothesis test, Linear Regression Analysis.

I. INTRODUCTION

Now the data is everywhere. The challenge is that how to work with data, how to have the analysis steps in the right flow i.e., where to start and how effectively carrying out the analysis and how to drive the meaningful inferences from the data which will really add value to the business.

This article explains how to start the analytical process right from data collection. Before starting the data collection, we need to have a clear understanding of the research problem and the result in which we are focused in. i.e., for what question we are expecting the answer. Then it will also address at what method to be used for data collection and what are the questions to be asked in case of survey method and how to frame the questions without any bios, How to fix the sample size etc.,

Before directly getting into Regression analysis, it is mandatory to do the exploratory data analysis by visualizing the data in to graphs to understand the relationship between the variables. The best fit line or linear regression line is the one that best describes the relationship between two variables. By looking into the graph, there is a possibility of identifying the patterns and then summary statistics to be calculated.

Once the Data Collection is over and we are ready with the data, we need to define the hypothesis for testing. Then the Single Variable Linear Regression may be used for predicting the inferences. These are explained in the subsequent topics.

Manuscript published on November 30, 2019.
* Correspondence Author

S. Yamini*, School of Computer Studies, Rathnavel Subramaniam College of Arts & Science (Autonomous), Sulur, Coimbatore, India. Email: yamini@rvsgroup.com

II. METHODOLOGIES USED

A. Describing and Summarising Data

Graphs are very useful in examining the data sets which is used to find trends and patterns and also helps in detecting the outliers. So the **Initial exploratory stage** is so valuable. Sometimes Descriptive statistics will be very useful since it is providing quick overview of data without mentioning every data point. The Correlation Coefficient is used to quantitatively measure the strength of the linear relationship between two variables on a scale from -1 to +1.

B. Sampling and Estimation

Due to time and resource constraints, it is not practically possible to always collect data from the entire population. So the representative random samples are collected, analysed and draw conclusion about the population.

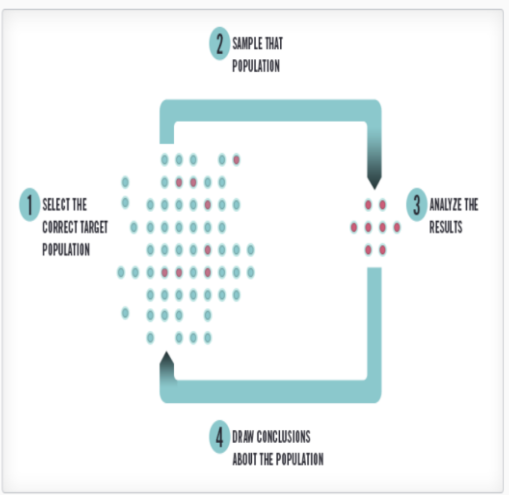

1 SELECT THE CORRECT TARGET POPULATION
2 SAMPLE THAT POPULATION
3 ANALYZE THE RESULTS
4 DRAW CONCLUSIONS ABOUT THE POPULATION

Fig. 1.Analysis Process.

How Sampling is used in Amazon:

Amazon is the most customer centric company. They are obsessed with the customer experience. To have the better customer experience, to improve the delivery process and to reduce the errors in the inventory, they use statistics and analytics a lot. They have many warehouses. The idea behind sampling in that situation is that randomly they pick a sample as guided by the software tool and check for the defects. This is the low cost way of doing the audit process in the inventory of the warehouses.

Samples may be collected in one of the three different ways.

- **Survey**: Researcher may directly ask the question from a random sample. Self-reported answers will be collected and analysed.
- **Observational Study**: Without intervening the process, just to allow it to naturally happen, the researcher will observe and collect data about how and what is happening.

Experimental Study: The Sample will be divided into 2 groups. One is called “Control Group” which is not manipulated. Another is called “Treatment Group” which is manipulated in the sense that the new idea or the method is implemented and the response is collected and compared with the control group.

Sometimes, the organizations use the combination of methods to collect the data. Amazon is frequently using A/B testing which is one of the Experimental Study methods for web design optimization. During the survey method, the researcher has to make sure that the type of question is unbiased and there must be high response rate.

Normal Distribution:

Once the sample is collected, it will be analysed to draw inferences about the entire population. For doing this, it will be useful to understand about the basic characteristics of common probability distribution called Normal Distribution.

Normal Distribution is the probability distribution that is centred at its mean. It is also referred as bell curve. It is shown in two axes. In x-axis, it shows the variable which we are interested in. y-axis is the likelihood of values of the variable which is used. As per the rules of thumb of the normal distribution, 68% of values will be lying within + or – one standard deviation distance. 95% of the values will be lying within + or – two (1.96 exactly) standard deviation distance and 99% of the values will be within + or – three standard deviation distance.

Standard Normal Curve is the Normal Distribution with mean as 0 and Standard Deviation as 1. The probability of values less than or equal to a particular value is called **Cumulative Probability**. It is conceptually related to the percentile of the distribution.

Central Limit Theorem gives much deeper insight about how sampling works. It says that if we take many random samples from the population and plot the means of each sample, assuming if we take large samples, then the resulting graph will be normally distributed invariably about the shape of the original population. The mean of the distribution means will be equal to the true population mean. But both distributions will be having different standard deviations. An important thing to be noted is that, the sample mean falls somewhere in the normal distribution which is centred at the true population mean. Because of this, we may completely disregard the distribution of the population and can concentrate only on samples.

Confidence Interval

Sample mean is only the point estimate. To make decisions as managers, we need more than a point estimate. We needs to know how close is our point estimate to the true population mean. For that we need to construct the confidence interval based on the sample mean and the specified level of confidence. That range is very likely to contain the true population mean. For example for 95% Confidence level, for 95% of all random samples, a range that is two standard deviations wide and cantered at sample mean contains the true population mean.

Often the people misinterpret and say that there is 95% probability or chance that the confidence interval contains the true population mean. This is completely wrong. We take a single sample and construct the confidence interval, either the range contains or do not contain the true population mean. There is no chance in that. Because we don’t know the true population mean, we cannot say that whether the particular samples confidence interval contains the true population mean or not. However we do know that on average, 95 out of 100 such intervals do contain the true population mean.

Amazon’s inventory Sampling

In one of the warehouse, Amazon used the cardboard dividers in between the shelf in the inventory to save money. They installed these dividers for the whole portion of the specific warehouse. After the couple of months, they find out that there is an increase in the defects in this particular part of the warehouse. They did the detailed analytics to check the reason behind this. Then they discovered that while picking the items, the pickers were bumping the items which flatten the cardboard and the items in the shelves were drifted. It was discovered through sampling.

C. Hypothesis Testing

To ensure that the important managerial decisions are well informed as possible, before making the decisions, the claims and theories need to be tested. Hypothesis testing allows to do this testing rigorously. Before starting the testing, null and alternate hypothesis to be clearly defined [2].

Null Hypothesis (H_0): It is the statement about our topic of interest. It is based on historical information or conventional wisdom. We always start the testing by assuming that null hypothesis is true and test to see if we can nullify that. It is the opposite of hypothesis we are trying to prove (Alternate Hypothesis).

Alternate Hypothesis (H_a): It is the claim or theory we are trying to substantiate.

There can be only two possible outcomes. We either Reject the Null Hypothesis or Fail to Reject the Null Hypothesis if we have insufficient evidence.

The outcome is based on the range of likely sample means. If the sample mean is outside this interval, then we can reject the Null Hypothesis. If the mean is within this interval then fail to reject the Null Hypothesis due to insufficient evidence against the Null Hypothesis.

P-Value: It is the Quantitative measurement of the likelihood of being the Null Hypothesis is true. If the p-value is less than the significance level, then we will reject the null hypothesis. If the p-value is greater than or equal to significance level then Fail to reject (Do not Reject) the Null Hypothesis.

Significance Level = 1 – Confidence Level

It is always important to use the managerial judgement when making decisions, especially when the p-value is very close to the significance level.

Type I and Type II Errors

Since the Hypothesis testing is based on Sample Data, there is always a possibility of drawing the wrong conclusion about the population. When we incorrectly reject the Null Hypothesis when it is actually true then it is Type I Error. If we incorrectly fail to reject the Null Hypothesis when it is not true then it is Type II Error.

The possibilities of the Type I Error is equal to the Significance Level and the since there is no information about the probabilities of different sample means, we cannot calculate the likelihood of Type II error.

D. Single Variable Linear Regression

Linear Regression is used to identify the linear relationship between the variables. It allows us to gain insights into the structure of the relationship and provides measures of how well the data fits in that relationship. It will be very useful in analyzing the historical trends and developing the forecasts.

The Regression line is the one that reduces the dispersion of the points around the line. The accuracy of the regression line is measured by that dispersion.

Forecasting

Once the linear Regression line is drawn, then the point forecast is possible using the line equation. The forecasting is uncertain if the independent variable is ranging outside the historical range of data. So it will be good to have an interval (Prediction Interval) instead of a single point which will increase the possibilities of the outcome to fall in that range. It based on the Confidence Level. The centre point of the interval is the Point Forecast.

Interpreting the Regression Output

Most of the time, our real time relationship may not be linear. In that case, it will be good to check whether the linear model is the best fit for the data. So we need to measure how helpful the linear model is in predicting the relationship between independent variable and dependent variable.

For this, first we need to measure the vertical distance between the data point and the regression line. This is called the residual error. To avoid the negative and positive values, these differences are squared and then added together that is called **Residual Sum of Squares**. This is the error which is **unexplained** by the regression Line. Suppose if independent variable's data value is not available, then the best predictor is the mean value of the dependent variable. So the difference between the data value and this line is squared and added up that is called **Total sum of Squares**. Now we compare these two values to better understand how much errors the Regression line is reduced. The difference between the Total sum of squares and the Residual sum of squares is called the **Regression Sum of Squares** (how much variation in dependent variable is explained by the independent variable) which are the errors **explained** by the Regression Line [9].

R^2 measures how closely the regression line that fits the data. It is the percentage of variation explained in the dependent variable by the independent variable. R^2 value ranges between 0 and 1. Just by looking in to the R^2 value alone, it is difficult to determine whether it is good or bad score since based on the applications and the scenarios, the prediction power will vary [10].

Testing for a Significant Relationship

We need to look beyond R2 to check how is the relationship between the two variable and whether the linear regression model is the best fit for the data. P-value of the independent variable and the residual plots also to be observed. If the P-value is less than the significance level and the Confidence interval is not having zero then there is a significant linear relationship exists.

Residual Plot

First we need to measure the vertical distance between the observed data point and the regression line which is the residual error. Then these measured values to be plotted against the independent variable in the graph which is called the **residual plot**. There should not be any systematic pattern in the residual plot if there is a linear relationship between dependent variable and the independent variable. The residuals should be spread randomly above and below the horizontal axis. Based on the assumptions of the linear regression, the residuals should form a normal distribution with mean zero and the fixed variance. If there is any systematic pattern between the residuals then the linear model may not be the best fit for this data.

For example, if the residuals are in the curved shape, there may be a nonlinear relationship exists. If the residuals become larger/smaller during the movement along x axis, then there may be a **heteroskedasticity** present in the data set.

III. CASE STUDIES

Amazon's use of Hypothesis Testing

The main focus of the Amazon is to be the most customer centric company in the universe. They are obsessed with the Customer Experience. To give best customer experience, they keep on updating their web and mobile application. But before implementing these changes, they do A/B testing [3][4]. Based on the outcome of the test, they either continue in implementing the changes to all the customers or they revert back if there is no significant impact.

Shopping cart A/B test

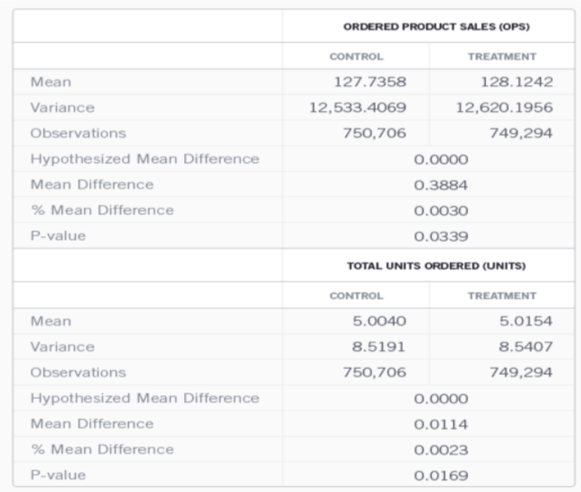
The amazon authorities' want to test whether displaying the number of items purchased in top of the shopping cart icon improve their sales. So they performed Hypothesis testing for that. They defined Control and Treatment Groups [5][6]. In the control group, there is no change in the existing website. For the treatment group, they incorporated the changes in the shopping cart icon which displayed the no of items purchased. They started the test by defining the null and Alternate Hypothesis.

H_0 : $\mu_{control} = \mu_{treatment}$; There is no change in the average money spent by the customers due to the change in the shopping cart.

H_a : $\mu_{control} \neq \mu_{treatment;}$ There is a change in the average money spent by the Customers due to the impact of the changes in the shopping cart design.

The sample size was 1.5 million observations.

Table- I: Result of Hypothesis Testing



	ORDERED PRODUCT SALES (OPS)	
	CONTROL	TREATMENT
Mean	127.7358	128.1242
Variance	12,533.4069	12,620.1956
Observations	750,706	749,294
Hypothesized Mean Difference	0.0000	
Mean Difference	0.3884	
% Mean Difference	0.0030	
P-value	0.0339	
	TOTAL UNITS ORDERED (UNITS)	
	CONTROL	TREATMENT
Mean	5.0040	5.0154
Variance	8.5191	8.5407
Observations	750,706	749,294
Hypothesized Mean Difference	0.0000	
Mean Difference	0.0114	
% Mean Difference	0.0023	
P-value	0.0169	

The result of this A/B test [7][8] for both the dollar sales and the units purchased were significant [1]. This has a huge impact on Customer Experience and improved the sales by 0.3% worldwide.

Regression at Disney Studios

In the movie Production Company like Disney Studios, there is a lot of distribution landscape like theatrical window, home entertainment window, online digital subscriptions etc., So the entire Return of Investment (ROI) is not recouped in only one window. The theatrical window is the most important one. Lot money is spent in marketing that window. Since this sets the path for the rest of the windows. It is really important for the Disney Studios to forecast about the home entertainment sales since that is one of the main parts of ROI and directly have a huge impact based on theatrical window.

In the Disney studios, fundamentally the forecasting of the home video units is based on the regression analysis of home video unit sales to the gross box office. It will go for three to four months of gross box office time to get the final value. Then what they will predict is the estimate of next 52 weeks of sales of home video units in the retail.

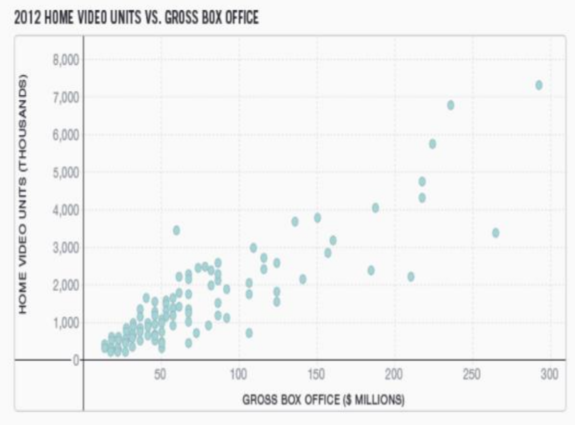


Fig. 2.ScatterPlot.

The above scatter graph shows the 2012 data of Home Video units sales and gross box office of Disney Studios [1]. It shows the strong relationship between the dependent and independent variables.

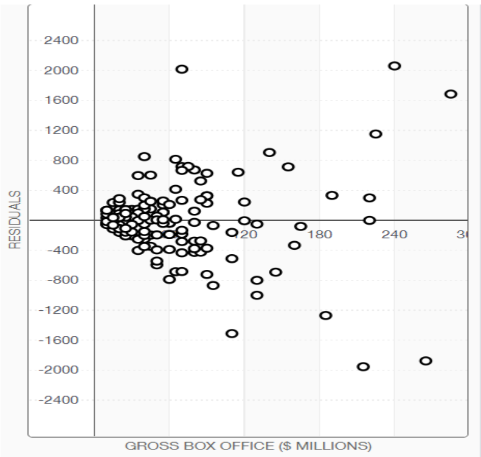
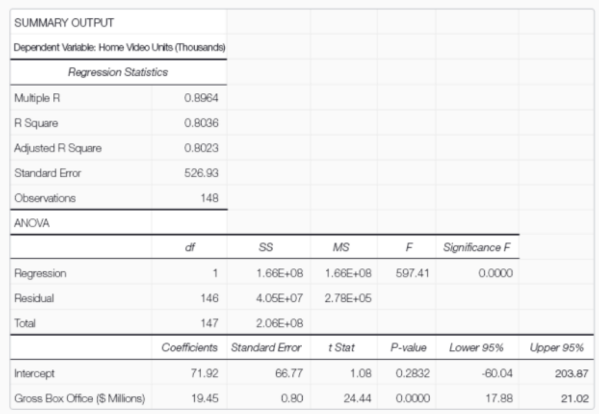


Fig. 3.Residual Plot.

The above graph is the residual plot of the Disney studio’s 2012 home videos units and the gross box office regression model [1]. The residuals are spread across the horizontal axis which shows that the linear model is the best fit for this data.

The regression output table given below is the output of the Regression Analysis of 2012 Home video units sales data [1]. It has high R^2 value of 80%. The p value is 0.0000 and the 95% confidence interval is not having the zero in it. So there is a significant relationship between the Home Video Unit Sales and the Gross Box Office Sales.

Table- II: Regression Analysis Output



SUMMARY OUTPUT						
Dependent Variable: Home Video Units (Thousands)						
Regression Statistics						
Multiple R	0.8964					
R Square	0.8036					
Adjusted R Square	0.8023					
Standard Error	526.93					
Observations	148					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	1.66E+08	1.66E+08	597.41	0.0000	
Residual	146	4.05E+07	2.78E+05			
Total	147	2.06E+08				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	71.92	66.77	1.08	0.2832	-60.04	203.87
Gross Box Office ($ Millions)	19.45	0.80	24.44	0.0000	17.88	21.02

Using the Single Variable linear regression is the best starting point for forecasting of the units because there is a strong correlation between the two variables. So R^2 in this case is high and it is a very good predictor. It may not be perfect but it is the good point to start the forecasting and the analysis.

IV. CONCLUSION

Due to the innovations in the technology, economic crisis and other factors, there is a tremendous shift in the consumer behaviors and it is happening quickly. It has complicated the Disney studios to accurately forecast the home video units. The other variables are also impacting the home entertainment beyond the box office like seasonality, gift giving events, trends in the industry etc.,

In today’s scenario, almost in all the industries the data is playing the major role in the business development as well as for making important managerial decisions.

So the organizations must have a strong team for doing Business Analytics with diverse skillsets.

During the analysis process, the team with different perspectives have to sit together, openly have to have a discussion then & there and come out with a solution or recommendation. Because the industry is having the different dynamics that is changing so rapidly, the point of view is so important. The data is certainly very important but the qualitative overlay is very important as well.

REFERENCES

1. Janice Hammond, “Business Analytics”, Harvard Business School, Harvard University, USA.
2. Ronny Kohavi, Thomas Crook, Roger Longbotham ,“ Online Experimentation at Microsoft”,2009.
3. Denis andrasec, Marcus bloice, Georg lexer, Jürgen zernig, “A/B Testing Literature Review”, Dec 2011.
4. “A/B Testing at Vungle”, Darden Business Publishing, University of Virginia, Mar 2017
5. Aaron Chatterji, “Experimentation and Startup Performance: Evidence from A/B Testing”, Rembrand Koning, Sharique Hasan, Harvard Business School, Aug 2019.
6. https://link.springer.com/referenceworkentry/10.1007%2F978-1-4899-7687-1_891.
7. https://www.forbes.com/sites/forbescommunicationscouncil/2017/04/05/how-ab-testing-can-drive-business-decisions/#3b4ae5977940
8. https://hbr.org/2017/09/the-surprising-power-of-online-experiments
9. Gohar Feroz Khan, Sokha Vong, “Virality over YouTube: an empirical analysis”, The University of Waikato,Sep 2014.
10. Iman Barjasteh, Ying Liu, Hayder Radha, “Trending Videos: Measurement and Analysis”, Michigan State University, Sep 2014.

AUTHORS PROFILE

Dr. S. Yamini is the Associate Professor in Computer Science and the Director (Academic), School of Computer Studies at Rathnavel Subramaniam College of Arts & Science (Autonomous), Sulur, Coimbatore, Tamil nadu, India. She is certified in Business Analytics by Harvard Business School, Harvard University, Boston, Massachusetts, US. Her primary interest and expertise is in Cyber Security and her research focuses on Multi level Network Security Design and Graphical Authentication System. She Completed the Micro Masters Certification Programme in Cyber Security which is offered by RIT (Rochester Institute of Technology) , New York in collaboration through edX platform and the Cyber Security specialization Programme which was offered by University of Maryland, USA through Coursera. She has also completed “Mathematics for Machine Learning” Course which is offered by Imperial College, London in the Coursera platform.