**Business research data analysis report.**

A picture containing graphical user interface

Description automatically generated

Aritha Jayaratne – s4683873

**Contents.**

[Introduction. 2](#_Toc119492341)

[Research Question and Objective. 2](#_Toc119492342)

[Research question. 2](#_Toc119492343)

[Research objectives. 2](#_Toc119492344)

[Variables. 3](#_Toc119492345)

[Hypothesis statements for testing. 4](#_Toc119492346)

[Research objective 2 – Region. 4](#_Toc119492347)

[Research objective 3 – Category 4](#_Toc119492348)

[Research objective 4 – Segment. 4](#_Toc119492349)

[Research objective 5 – Ship mode. 4](#_Toc119492350)

[Methodology. 5](#_Toc119492351)

[Results, Interpretations, Findings and Discussions. 6](#_Toc119492352)

[Data cleaning. 6](#_Toc119492353)

[Data Coding. 7](#_Toc119492354)

[Data analysis. 8](#_Toc119492355)

[Descriptive analysis of variables. (Research Objective 1) 8](#_Toc119492356)

[Statistical analysis of variables (region, category, ship mode & segment) 13](#_Toc119492357)

[Recommendations. 17](#_Toc119492358)

[Limitations of this study. 18](#_Toc119492359)

[Conclusion. 19](#_Toc119492360)

[References. 19](#_Toc119492361)

# **Introduction.**

Data driven decision making has developed a lot in the 21st century. Multibillion dollar companies focus and spends a huge amount of their budgets on analyzing various types of data to make business decisions. The above research was conducted on a data set which was relating to the Amazon online retail company. The research was conducted on Amazon, online and retail giant based in United states to find out it’s customer basis, shipping mode the customers use, and which region of United States of America has more sales and profits. Various analytical tests were performed to derive the outcomes and recommendations to the company to get a better understanding.

# **Research Question and Objective.**

## Research question.

Question 1 – What is the descriptive analysis of all the variables of the collected data relating to Amazon?

Question 2 – How does Amazon compare the sales of east and west region?

Question 3 - What is the most performing category out of furniture and technology which has an outstanding performance in sales in the Amazon company in United states?

Question 4 – What is the difference of sales in the three consumer segments of Amazon in United states?

Question 5 – What shipping mode does the highest number of customers use out of the 4 categories to ship their products according to sales on Amazon?

## Research objectives.

Objective 1– To understand the descriptive analytics of the given data set.

Objective 2 - To analyze the regional sales performance based on its sales.

Objective 3 - To understand the product category’s performances and compare them according to its sales.

Objective 4 - To analyze the consumer segments based on the number of sales made by each segment.

Objective 5 - To compare which shipping mode does the most and the least number of sales.

## Variables.

In the chosen data set the primary variable has been marked in green. The grouping variables which have only 2 categories have been marked by blue color in the excel sheet. And the grouping variables which have more than 2 categories are marked by yellow on the excel sheet.

Primary variable:

* Sales

Grouping variables (2 categories): Region

* West
* East

Grouping variables (2 categories): Category

* Furniture
* Technology

Grouping variables (3 categories): Segment

* Consumer
* Corporate
* Home office

Grouping variables (4 categories): Ship mode

* First class
* Same day
* Second class
* Standard class

# **Hypothesis statements for testing.**

## Research objective 2 – Region.

Null Hypothesis (Ho): The population distribution of sales of the retail supermarket in West and East regions of United States is symmetric around 0.

Alternate Hypothesis (H1): The population distribution of sales of the retail supermarket in West and East regions of United States is not symmetric around 0.

## Research objective 3 – Category

Null Hypothesis (Ho): The population distribution of sales of the retail supermarket in furniture and Technology categories is symmetric around 0.

Alternate Hypothesis (H1): The population distribution of sales of the retail supermarket in furniture and Technology categories is not symmetric around 0.

## Research objective 4 – Segment.

Null Hypothesis (Ho): Sales means are equal in each consumer segment.

Alternate Hypothesis (H1): Sales means are not equal in each consumer segment.

## Research objective 5 – Ship mode.

Null Hypothesis (Ho): Sales means are equal in each ship mode.

Alternate Hypothesis (H1): Sales means are not equal in each ship mode.

# **Methodology.**

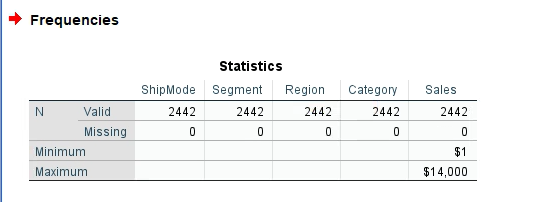
The competition is increasing in the retail business in United states of America. Many companies are trying to make the correct decision for their company to avoid loss of sales and to make decisions which will gradually increase their sales. The data set was published on the Kaggle website, and this was used to analyze the secondary data for this report relating to the online sales of Amazon. In the data set there are various ways which the author has used to collect the data. Variables such as order Id, customer id, segment, category, ship mode and subcategory to collect the data. There is a total of 2442 participants. In the selected data set, most of the variables were descriptive. To perform a statistical and descriptive analysis the data had to be cleaned. Two steps were used in the cleaning process. The first step was the missing value analysis and the other was finding the z core of the variables and detecting the outliers and removing them. Due to the data set mostly being descriptive a re-coding was performed on the variables so that it could be analyzed. The data cleaning steps, and the re-coding are described in detail in the results, interpretations, findings, and discussions section.

For the statistical analysis of the data the team used SPSS software. A analysis was performed on the data set to see if it’s normally distributed or not normally distributed. To conclude the descriptive analysis part, research was conducted on the factors such as Mean and Median, Shape of the Histogram, skewness, and Kurtosis. Regarding the region and the category, a Wilcoxon rank sum test was performed on the data set because it was not normally distributed. And for Segment and ship mode a decision was made to perform the Kurksal Wallis test because the data is not normally distributed in both the data sets. Another reason for choosing the Kurksal Wallis test is because the data set is extremely large which contained more than 2000 variables and it is not normally distributed. Other tests such as T test, Paired t test and anova test was not conducted due to the reason, the data set being not normally distributed.

# **Results, Interpretations, Findings and Discussions.**

## Data cleaning.

Step 1 – Missing value analysis.



After doing a manual data cleaning session and analyzing the data set through SPSS to find any missing data. After running the data set through SPSS and analyzing the main variables for missing data that will be used to perform the data analyzing tests, above table shows the statistics of ship mode, segment, region, category, and sales with zero missing variables.

Step 2 – Detecting and removing outliers.

Table

Description automatically generatedAn outlier testing was performed on the data set which included the testing variables. Only the sales variable had positive outliers which were larger than +3.0. Other variables didn’t have any outliers in the data set. After removing the outliers, the data set was checked again for the normality of the distribution. There was a difference in the sales variable but there was no big difference in the other variables. After clearing the outliers, the no of data was decreased from 2442 to 2400.

Data set before removing outliers.

Data set after removing outliers.

## Data Coding.

In this report all the variables of region, category, segment, and ship mode has been recoded and given numerical values to be used for analyzing and statistical interpretations. The data has been coded as follows and as proof a SPSS file will be submitted containing the coding.

Region: Region recoded as “Region\_code”.

“West – 1”, “East – 2”

Category: Category recoded as “Category\_code”

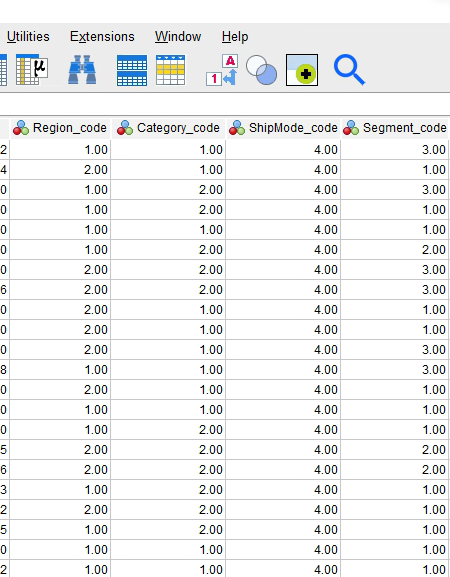
“Furniture – 1”, “Technology – 2’’

Segment: Segment recoded as ‘’Segment\_code’’

“Consumer – 1”, ‘’Corporate – 2”, ‘’Home office – 3’’

Ship mode: Ship mode recoded as ‘’Ship mode\_code’’

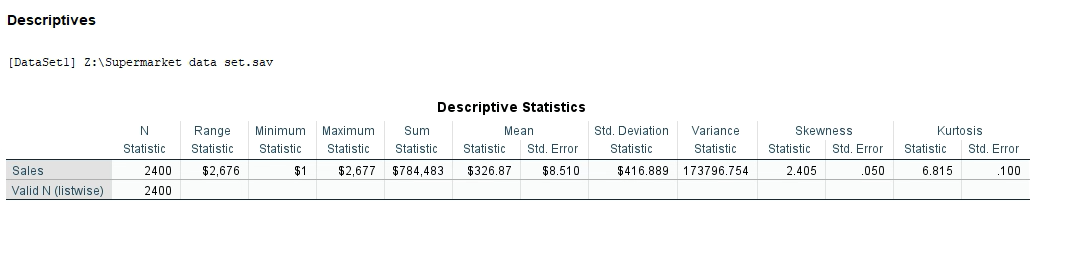
‘’First class – 1’’, ‘’Same day – 2’’, ‘’Second class – 3’’, ‘’Standard class – 4’’

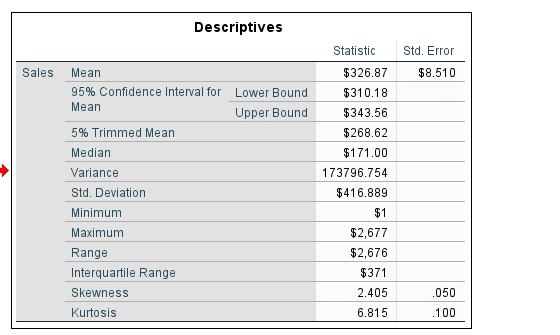


## Data analysis.

### Descriptive analysis of variables. (Research Objective 1)

1. Sales.



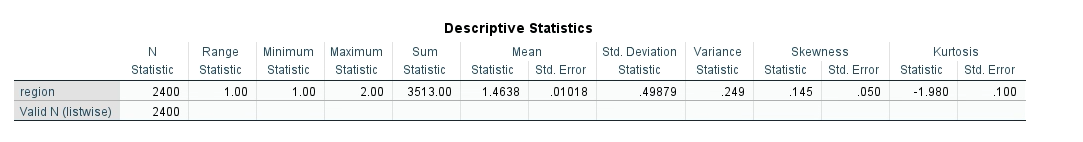


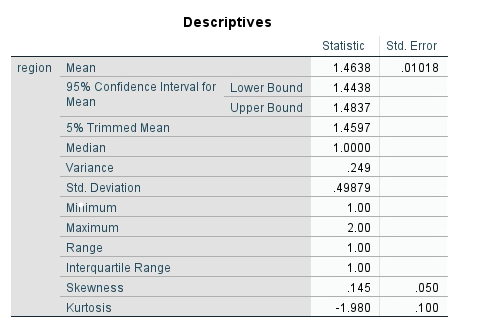
Chart, histogram

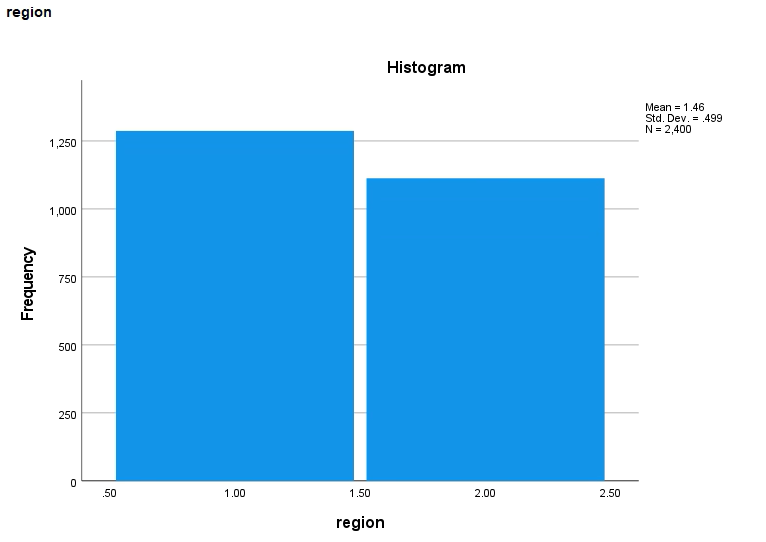
Description automatically generated

The first variable which was analyzed was the primary variable “sales”. The total number of variables in the sales column is 2400 and the mean is $326.87. The minimum value of this data set is $1, and the maximum value is $2,677. According to the data analyzed there. The skewness is 2.405 and the kurtosis is 6.815. The data set of the sales variable does not have a normal distribution. Reason for this conclusion is the shape of the histogram in the above picture is positively skewed and the mean is greater than the median (326.87 > 171.00). Kurtosis also has a positive value, and the skewness is also positive. The data set has a leptokurtic distribution. Due to these reasons the data set of the sales variable is not normally distributed.

1. Region.



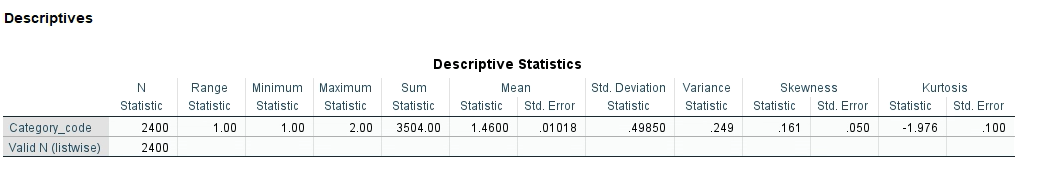


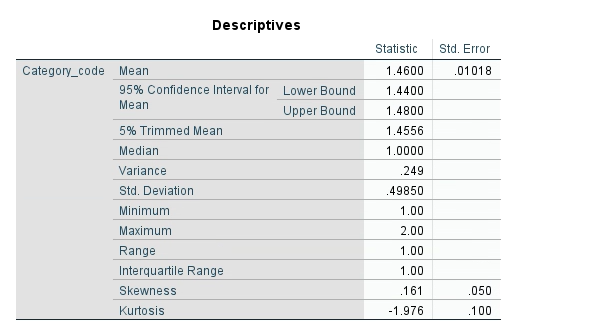
Chart, pie chart

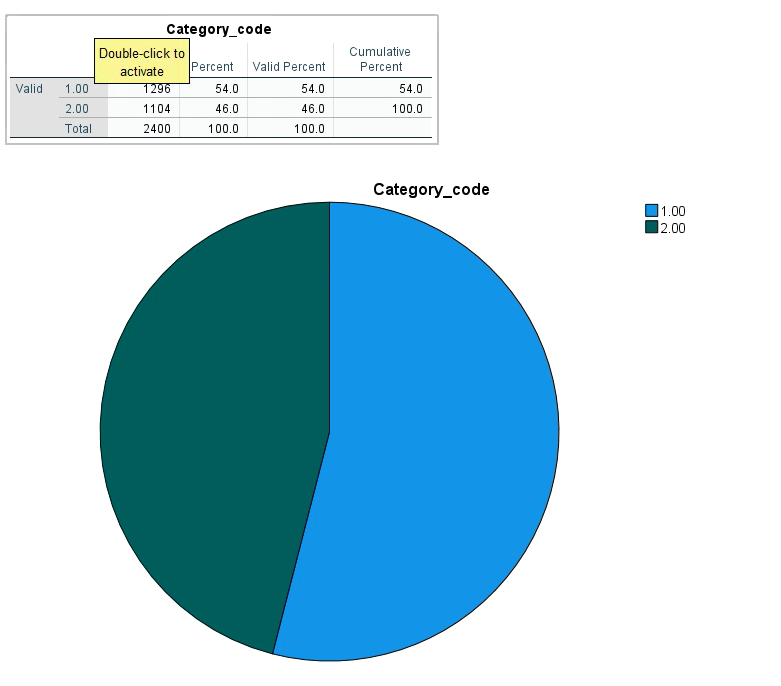
Description automatically generated

After recording the region column into numerical values, the above analysis was conducted regarding the data set. Region variable is divided into 2 categories as West and East. The number of variables available in this data set is N = 2400. The mean of this data set is 1.46 and the median is 1. As per the above histogram the data set is positively skewed, and the mean is greater than the median (1.46>1). Due to these observations, it is safe to say that the data is not normally distributed. Due to this reason the team decided to conduct the Wilcoxon rank sum test on the two variable data set.

1. Category.



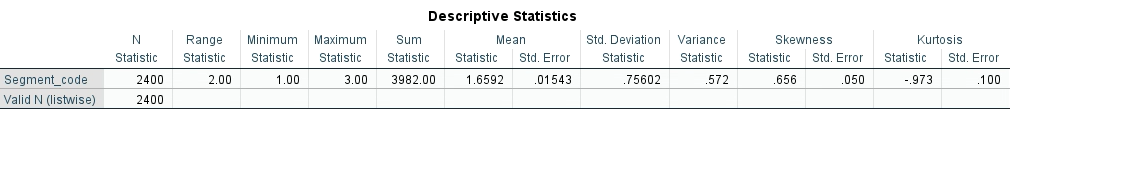


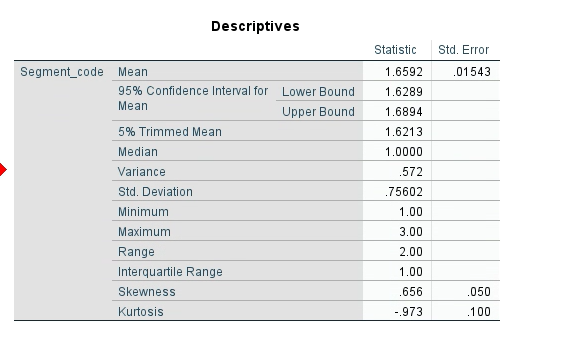
Chart, bar chart, histogram

Description automatically generated

After recoding the category data set a descriptive analysis was performed using SPSS software to check the distribution of the data set to perform the relevant statistical tests. The first factor which our team investigated was the mean and the median comparison. The mean is 1.46 and the median is 1. The mean is greater than the median (1.46>1). Due to this the team concluded that the data set is not normally distributed. The other way was the shape of the histogram. The data set was more skewed to the right and the kurtosis had a negative value (-1.976). The distribution is more platykurtic. The conclusion was that the data is not normally distributed and therefore a Wilcoxon rank sum test will be performed.

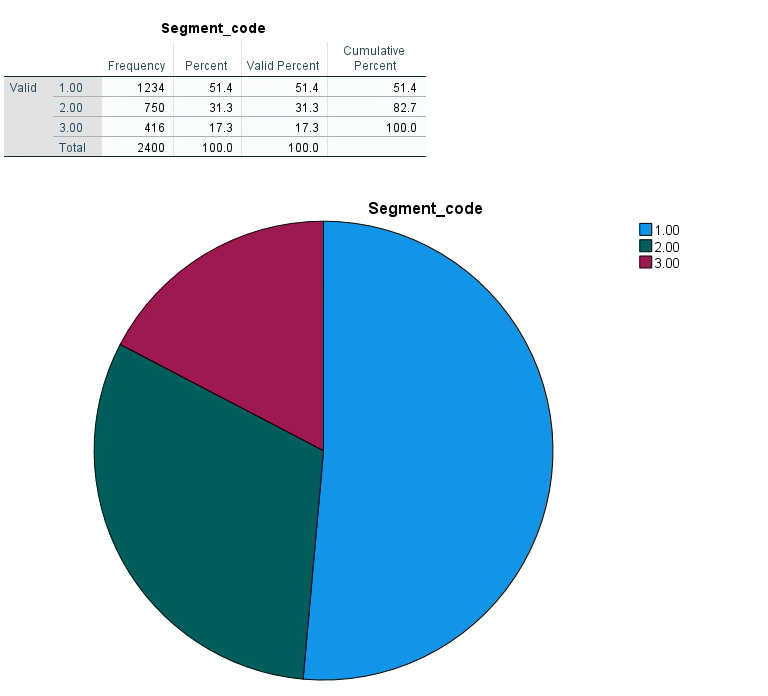
1. Segment.





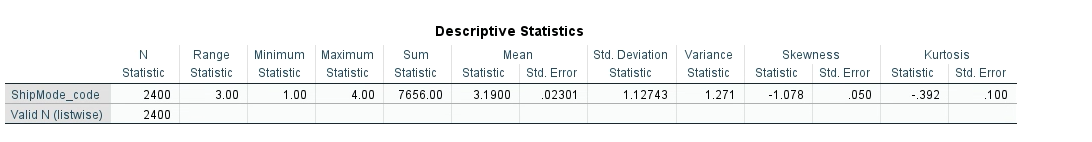
Chart, histogram

Description automatically generated



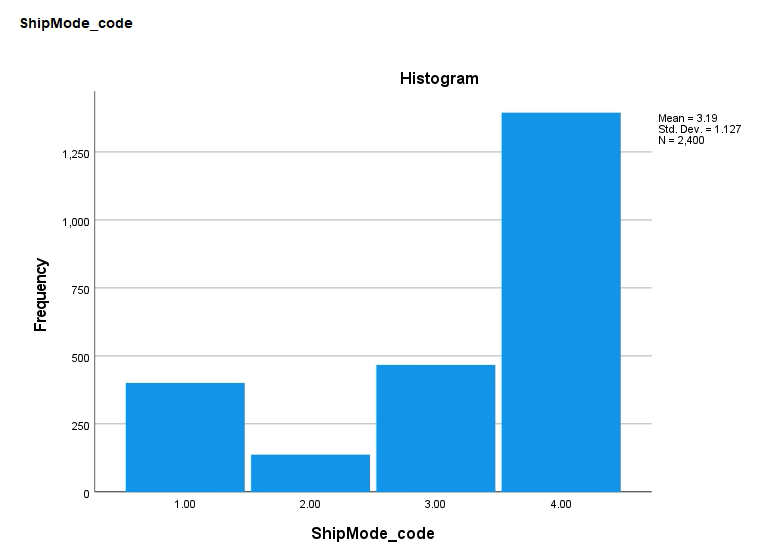
After recoding the segment data set a descriptive analysis was done and by using the above tables, a decision was made regarding the distribution of the data set. The first way was to compare the mean and the median. The mean is 1.66 and the median is 1. The mean is greater than the median (1.66>1). The other factor which was considered was the shape of the histogram. The histogram is skewed to the right and has a negative kurtosis which is -.973. After analyzing this data, it is safe to say that the data is not normally distributed so the team will perform the Kruskal Wallis test.

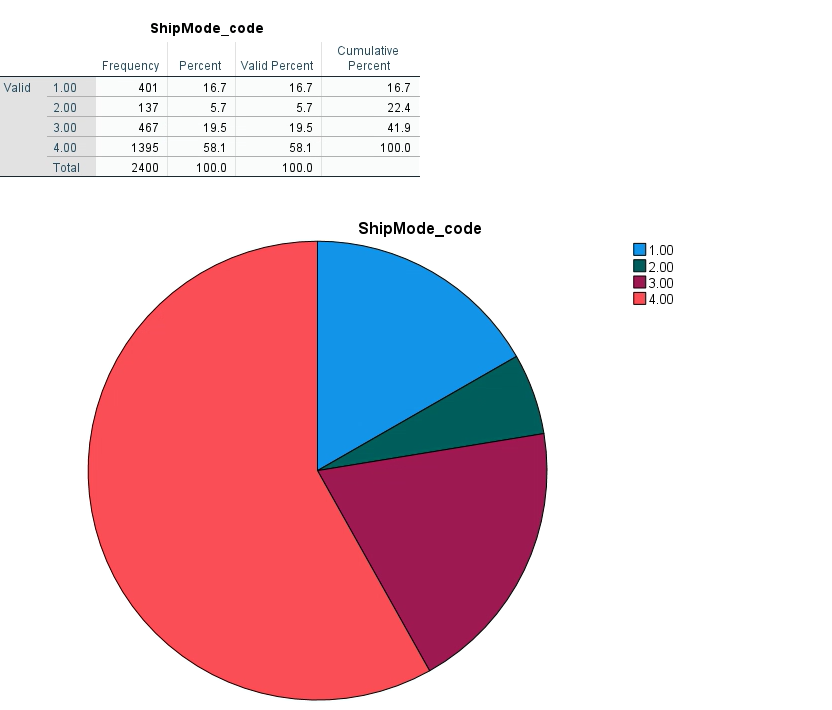
1. Ship mode.



Table

Description automatically generated



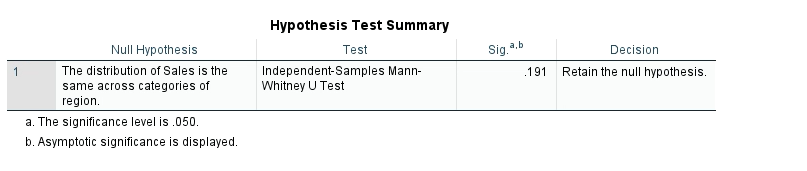


The ship mode category was also recoded to perform the descriptive analysis using SPSS. This was done to analyze the data set to see if it’s normally distributed or skewed. The mean of this data is 3.19 and the median is 4. In this data set the median is greater than the mean (4>3.1900). There for the data is left skewed and it is not normal. The shape of the histogram is also skewed to the left and it also has a negative kurtosis value which is platykurtic distribution. Therefor the test that will be conducted is the Kruskal Wallis test.

### Statistical analysis of variables (region, category, ship mode & segment)

#### Research objective 2 - Region (Wilcoxon rank sum test.)

After conducting tests to check the normality of the data sets of region and sales, Wilcoxon rank sum test was performed and below mentioned statistical decisions were made. the significance level of 0.05% was used will performing the test.



Hypothesis.

Null Hypothesis (Ho): The population distribution of sales of the retail supermarket in West and East regions of United States is symmetric around 0.

Alternate Hypothesis (H1): The population distribution of sales of the retail supermarket in West and East regions of United States is not symmetric around 0.

Statistical decision.

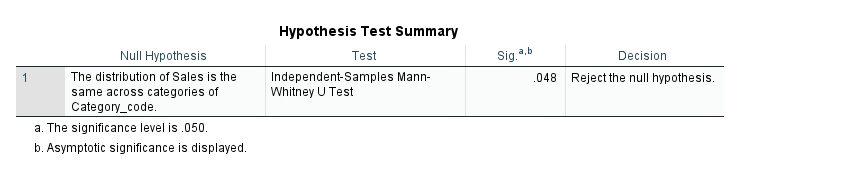
Accept the Ho because P value is greater than 0.05 (0.191>0.05).

Conclusion.

At 5% level of significance, the sample evidence supports the null hypothesis that the population distribution of differences in sales of the supermarket in the west and East regions of United states of America is symmetric.

#### Research objective 3 – Category (Wilcoxon rank sum test).

After the conclusion that the category data set was not normally distributed, the Wilcoxon rank sum test was performed at a significance level of 0.05%. The Statistical analysis is interpreted below.



Hypothesis.

Null Hypothesis (Ho): The population distribution of sales of the retail supermarket in furniture and Technology categories is symmetric around 0.

Alternate Hypothesis (H1): The population distribution of sales of the online retail store in furniture and Technology categories is not symmetric around 0.

Statistical decision.

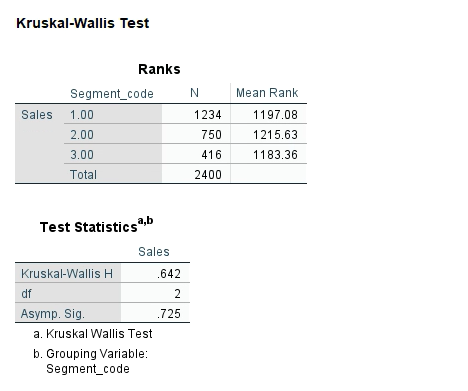
Accept the H1 because the P value is less than 0.05 (0.05>0.048).

Conclusion.

At 5% level of significance, the sample evidence does not support the null hypothesis that the population distribution is different in sales of the online store category of furniture and technology, and it is not symmetric.

#### Research objective 4 – Segment (Kruskal Wallis test).

After concluding that the segment data set was not normally distributed by analyzing the mean, skewness and kurtosis and the histogram shape. After a careful analysis the team decided to perform the Kruskal-Wallis’s test having the significance level at 0.05%. Statistical interpretation is mentioned below.



Hypothesis.

Null Hypothesis (Ho): Sales means are equal in each consumer segment.

Alternate Hypothesis (H1): Sales means are not equal in each consumer segment.

Statistical decision.

Accept the Ho because the P values is greater than 0.05 (0.725>0.05).

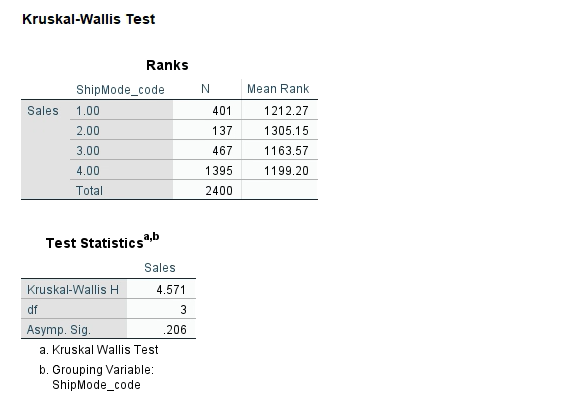
Conclusion.

At a significant level of 0.05 the distributions of mean are all the same in each consumer segment.

According to the statistics of the test performed the sale mean is equally distributed among all customer segments of the Amazon company.

#### Research objective 5 – Ship mode (Kruskal Wallis test)

After analyzing the ship mode data and referring to the statistical analysis it was decided that the data was not normally distributed because of its mean and median, histogram shape and the skewness and kurtosis. The Kruskal Wallis test will be conducted to support the hypothesis statement of research objective 4. The significance level was at 0.05%.



Hypothesis.

Null Hypothesis (Ho): Sales means are equal in each ship mode.

Alternate Hypothesis (H1): Sales means are not equal in each ship mode.

Statistical decision.

Accept the Ho because the P values is greater than 0.05 (0.206>0.05).

Conclusion.

At a significance level of 0.05 the distribution of mean is all the same in each Ship mode category.

According to the statistics of the test performed the sale mean is equally distributed among all shipping modes in the Amazon company.

# **Recommendations.**

As one of the best online retailing companies in the world, Amazon has a duty to provide the best service to the customer and at the same time make sales and turn it to profits. After running the tests as sales as the primary variable in the data set gathered. The sales variable was only used on four grouping categories from the collected the data.

The marketing teams of Amazon who oversee the east region should focus more to develop and do more marketing campaigns to increase the sales in the region. Mainly the company could focus on social media, giving discounts and coupons and plan more campaigns in the festive season.

Also, in the category variable the furniture category is making more sales compared to technology category. To balance this the company could cut their profit margin and pass it to the customer as a discount so that the customers are more tempted to buy the most discounted products when it comes to technology.

Another recommendation would be to increase the number of customers in the cooperate and the home office segments. A reason to this is that it would be easier to earn more sales from corporate sector as they will be doing bulk purchasing.

A recommendation for the Amazon company would be to reduce the prices or discounts to customers when using first class, same day and second class. Then they will e more tempted to try the other three shipping modes due to the discounts and loyalty points they will get.

These are the recommendations that have been recommended to the Amazon company based on the descriptive and statistical research conducted.

# **Limitations of this study.**

According to the data analysis of the data set conducted about this research this essay will address some limitations for this study.

The data analysis has choses different variables such as ship mode, consumer segment, region and category and the dependent variable of these three variables is the sales variable. More specific variables relating to this study would be gender, age. But the author of the data set has not gathered this information because he has used the customer id and the product Id in the data collection process. There for it is a bit hard to interpret the data more freely. Also, the statistical analysis does not depict the full idea of the test because when gathering data most of the data sometimes is ignored or cannot be used. There for It do not analyze the full conclusion or the result. In this study also some data had to be deleted before conducting the tests due to them being outliers of the data set. Due to this it is sometimes hard to analyze the exact results. Also, the data set which was used in this data set was a secondary data set, this was not gathered according to the needs of this report even though variables were help. If primary research was conducted to gain the necessary information needed the analyzing tests would have given out even more statistical decisions relating to report to make the necessary decisions.

# **Conclusion.**

In conclusion the secondary data set which was used in this report was very much helpful to gather important insights and make statistical business decisions based on the primary variable sales. Amazon company should maintain daily collection of data to make statistical decisions to maximize its sales and profit margins and concentrate on the needs of the customers. In conclusion the report was able to analyze many useful insights relating to the variables and recommendations were suggested to the company in how to improve or increase their sales.

# **References.**

* Lin, T. *et al.* (2021) ‘Extending the Mann-Whitney-Wilcoxon rank sum test to survey data for comparing mean ranks’, *Statistics in Medicine*, 40(7), pp. 1705–1717. doi:10.1002/sim.8865.
* Rosner, B, Glynn, RJ & Lee, M-LT 2003, ‘Incorporation of Clustering Effects for the Wilcoxon Rank Sum Test: A Large-Sample Approach’, *Biometrics*, vol. 59, no. 4, pp. 1089–1098.
* Johnson, R.W. (2022) ‘Alternate Forms of the One-Way ANOVA F and Kruskal-Wallis Test Statistics’, *Journal of Statistics Education*, 30(1), pp. 82–85. doi:10.1080/26939169.2021.2025177.
* Mahmoud Miari, Mohamad Taher Anan & Mohamed Bisher Zeina 2022, ‘Single Valued Neutrosophic Kruskal-Wallis and Mann Whitney Tests’, *Neutrosophic Sets and Systems*, vol. 51, pp. 948–957.