HOW DOES INCOME AFFECT MONTHLY CHARGES ON INTERNET SERVICES?

EXPLORATORY DATA ANALYSIS – D207

PERFORMANCE ASSESSMENT TASK

STUDENT NAME: IBRAHIM SULEIMAN

DATA ANALYTICS

STUDENT NUMBER: 001429984

COLLEGE OF INFORMATION TECHNOLOGY, WESTERN GOVERNORS UNIVERSITY

NAME OF INSTRUCTOR: DAVID GAGNER

May 23th, 2022

**Exploratory Data Analysis (EDA)** is the path to evaluating data to compile their paramount quality, recurrently alongside viewable techniques. However, we execute an analysis on data compiled to acquire essential metrics and features by applying part of the commendable and cordial representation.

**PART A. Describe a real-world organizational situation or issue in the Data Dictionary you chose, by doing the following:**

**1. Provide one question that is relevant to your chosen data set. You will answer this question later in the task through an analysis of the cleaned data, using one of the following techniques: chi-square, t-test, or analysis of variance (ANOVA).**

**The submission provides a specific question to be addressed through the analysis of the data set using one of the listed analysis techniques. The question is relevant to the chosen data set.**

As exemplified in the course, exploratory data analysis explores the essential qualities and performance paramount to executing parametric or non-parametric hypothesis examination to resolve the spreading of collected data, measure central tendency, and project the statistics visibly. The churn data set is elected for this performance assessment. The terminology Churn, in the contemporary business universe, could be described as the ratio of customers who eliminate their subscriptions that aren’t available for recommencement at the end of their subscription cycle. Similarly, churn depicts the number of customers who discontinued utilizing the brand type. However, I aim to execute an investigation on a churn dataset to create a solution to a contemporary business inquiry such as – how does income affect monthly charges on internet services?

A **t**-test is an analytical test also known to be used to contrast the means of dual formation. It is time after time passed down in hypothesis testing to decide a case a technique possesses implementation on the desired group set alongside if they appear to have diversity among themselves.

T-test will be used to compare the means of online security in relation to the monthly charge. This will help to determine whether these variables are different from one another that is normality, equal variance, and independence.

An Independent t-test will be used on online security and income. We will compare the mean of online security to a pre-specified value and test for a deviation from that value.

We could know that the average of online security would be given and specific and wish to compare the average number of customers who said online security(No) as not having online security.

***2.* Explain how stakeholders in the organization could benefit from an analysis of the data.**

***The explanation correctly addresses how stakeholders in the organization could benefit from the data analysis.***

The question applies to stakeholders in the organization because it reflects a solution to the issues and analyzes the fundamental value of shaping an inventive quick fix for the commitment of ownership of internet security positioned on the customer’s income and subscriptions with interest to the rate at which internet service is used. This will help in making key recommendations.

***3.* Identify all of the data in your data set that are relevant to answering your question in part A1.**

***The submission correctly identifies the specific data within the data set that are relevant to addressing the question from part A1.***

The identified variables accessible in a data set are as below:

**Income**: Annual income of customer (or invoiced person) as reported at the time of sign-up.

**MonthlyCharge:** The amount charged to the customer monthly. This value reflects an average per customer. For brand new customers, this value is the average for other customers who fit the new customer’s profile.

**OnlineSecurity;** Whether the customer has an online security add-on(Yes,No).

**InternetService:** Customer’s internet service provider (DSL, fiber optic, None).

**PART B. Describe the data analysis by doing the following:**

**1. Using one of the following techniques, write code (in either Python or R) to run the analysis of the data set:**

***The submission includes a warning- and error-free code to accurately analyze the data set using 1 of the given techniques.***

|  |
| --- |
| #lets perform t-test for multiple columns after groupby pandas to test the mwan of two different groups of multiple, columns in pandas  df = pd.DataFrame(np.random.randint(0,150,size=(100, 2)), columns=['col\_1', 'col\_2'])  df['OnlineSecurity'] = ['Yes']\*50 + ['No']\*50  df.groupby('OnlineSecurity').agg({"col\_1":"mean","col\_2":"mean"})  #we can iterate columns and perform more t-test by groups  #t\_col\_1 is the t statistics of the difference of means of col\_1 in the female, male and prefer not to say.  tstats = {}  ix\_a = df['OnlineSecurity'] == 'Yes'  for x in df:  if x != 'OnlineSecurity':  tstats['t\_' + x] = stats.ttest\_ind(df[x][ix\_a], df[x][~ix\_a])[0]    df.groupby('OnlineSecurity').mean().assign(\*\*tstats)  #H0: the means of the samples are equal  #H0: the means of the samples are unequal  #T-test for OnlineSecurity  from scipy.stats import ttest\_ind  No = [76.76, 76.42]  Yes = [80.16, 81.44]  stat, p = ttest\_ind(No, Yes)  print('stat=%.3f, p=%.3f' % (stat, p))  if p > 0.05:  print('Probably the same distribution')  else:  print('Probably different distributions') |

***2*. Provide the output and the results of any calculations from the analysis you performed.**

***The submission includes the output from running the code and the results of all calculations performed.***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | **col\_1** | **col\_2** | | --- | --- | --- | | **OnlineSecurity** |  |  | | **No** | 76.76 | 76.42 | | **Yes** | 80.16 | 81.44 |  |  | **col\_1** | **col\_2** | **t\_col\_1** | **t\_col\_2** | | --- | --- | --- | --- | --- | | **OnlineSecurity** |  |  |  |  | | **No** | 76.76 | 76.42 | 0.378115 | 0.561035 | | **Yes** | 80.16 | 81.44 | 0.378115 | 0.561035 |   stat=-6.358, p=0.024  Probably different distributions |

***3*. Justify why you chose this analysis technique.**

***The justification addresses why the chosen method of analysis was selected. The justified technique is one of the three listed in part B1. The technique is sufficient and appropriate for the chosen data set and addresses the question in part A1.***

The t-test is chosen for this analysis because it connotes a numerical test that is used to contrast the means of two groups. It is often used in hypothesis testing to regulate in case a mechanism or procedure has a consequence on the population of significance, as an alternative even if dual collections are distinctive against some other.

T-test entails two categories:

(a) dependent-means t-test which portrays equated pairs or duplicated measures t-test: This is applicable when the indistinguishable subjects partake in the pair conditions of the analysis.

(b) independent-means t-test same as the autonomous measures t-test: we apply this method in the scenario where it appears we have double diverse sets of subjects, a single set operating single condition in the analysis, alongside the different set operating the different condition.

Therefore, in this scenario, we get hold of a single independent variable (this is what we employ in the analysis process), alongside two levels (the two diverse conditions of our analysis). We possess a single dependent variable (the factor that we significantly calculate its dimension).

Therefore, the null hypothesis (H0) is that the true difference between these group means is zero in a given distribution.

On the other hand, the alternate hypothesis (Ha) reflects the true difference is different from zero in a given distribution

**PART C. Identify the distribution of two continuous variables and two categorical variables using univariate statistics from your cleaned and prepared data.**

***The submission accurately identifies the distribution of 2 continuous and 2 categorical variables using univariate statistics.***

1. ***Represent your findings in Part C, visually as part of your submission.***

The submission accurately represents the distribution of the variables in part C.

Univariate statistics entails a single dependent variable and can include one or more independent variables. Therefore, histograms and box plots are some of the most commonly used univariate statistics methods to represent the data visually.

From the selected datasets we will consider the following variables –

* **Continuous variables – Monthly Charge and Income**
* **Categorical variables - Internet service and online security**

**2 Continuous variables – Monthly Charge and Income**

|  |
| --- |
| df['Income'].hist()  <AxesSubplot:>    boxplot=sns.boxplot(x='Income',data=df)    df['MonthlyCharge'].hist()  <AxesSubplot:>    boxplot=sns.boxplot(x='MonthlyCharge',data=df) |

**2 Categorical variables - Internet service and online security**

|  |
| --- |
| InternetService= df.groupby(by="InternetService").size()  %matplotlib inline  InternetService.plot.bar()  <AxesSubplot:xlabel='InternetService'>    df['OnlineSecurity'].hist()  <AxesSubplot:> |

**PART D. Identify the distribution of two continuous variables and two categorical variables using bivariate statistics from your cleaned and prepared data.**

***The submission accurately identifies the distribution of 2 continuous and 2 categorical variables using bivariate statistics.***

1. ***Represent your findings in Part D, visually as part of your submission.***

**The submission accurately represents the distribution of the variables in part D.**

**The bivariate statistics, more than one variable is considered(two). Therefore, the variables below will be considered-**

To ascertain a vivid understanding of the variables, we are applying scatterplot and heatmap for bivariate analysis systematically to unveil the relationship between two continuous variables which would be measured on the ratio scales and intervals scales.

* **Continuous variables – Monthly Charge and Income**
* **Categorical variables - Internet service and online security**

**From the below scatter plot, it appears there is the absence of a strong relationship between the chosen variables.**

|  |
| --- |
| #lets check the relationship analysis, corelation matrix  corelation = df.corr()  sns.heatmap(corelation, xticklabels=corelation.columns, yticklabels=corelation.columns  ,annot=True)  <AxesSubplot:>    # Create a scatterplot of distribution of two categorical variables using bivariate statistics  sns.scatterplot(x='InternetService', y='OnlineSecurity', data=df)  # Show plot  plt.show() |

2 Continuous variables – Monthly Charges and Income

|  |
| --- |
| ***sns.scatterplot(x='MonthlyCharge', y='Income', data=df)***  ***# Show plot***  ***plt.show()*** |

**PART E. Summarize the implications of your data analysis by doing the following:**

1. ***Discuss the results of the hypothesis test.***

**The discussion includes accurate and complete results of the hypothesis test**.

Upon conducting a hypothesis test for a mean, The p-value of 0.024 is less than 0.05.

Since the p-value of 0.024 is less than the significance level of 0.05, then it rejects the null hypothesis and concludes that there is sufficient evidence to say that the true average of online security stands to be a factor that influences the original question, “there is a significant relationship between online security and monthly charge, as well as having diverse income rate that affects internet service?” because It's more accurate to say that you have 2.4% chance of incorrectly believing that the test scores are different (in a two-tailed test).

Therefore, It has the potential for adding key information to the assessment of income aimed at affecting the monthly charges of customers that use the internet service types(optic fiber, DSL, and others with none within these categories).

1. ***Discuss the limitations of your data analysis.***

**The explanation includes the limitations of the data analysis and does not include limitations that do not apply to the analysis.**

we could explore the hypothesis of the t-test, therefore the constraints appear to be unclouded. whereas data oppose the hypothesis, t-tests could appear to have fewer percentages of authenticity.

Therefore, these hypotheses encompass the following:

* The proportion of estimation. The expectation for a t-test is that the ratio of analysis enforced to the statistics acquired proceeds from a continuous or ordinal scale, equivalent to the amount for an IQ evaluation.
* The Univariate Imputation Technique also has certain limitations when it comes to the analysis because it could account for ambiguity and confusion as a result of lost data.
* An unplanned sampling is also known as random sampling. The statistics are acquired in distinction to a presentation, randomly elected quantity of the entire population.
* Data is commonly dispersed.
* The dualistic populations retain the comparable variance. This could therefore be transformed into an analysis.
* The dualistic populations retain the comparable variance. This could therefore be transformed into an analysis.

1. ***Recommend a course of action based on your results.***

**The recommendation includes *both* a response to the question from part A1 and specific actions that could be taken in response to the analysis. The recommendation is relevant to the situation and question and would plausibly address the situation and question**.

In distinction to this analysis, it is suggested to administer extra stability to online security concerning alternative variables being hypothesized, this could improve extra determination in the customers, and the churn rate could be scaled down as a result of the secured platform atmosphere of internet service activated by customers, despite their contradistinctive income matched together with subscription rating analytically measured in their monthly charges respectively.

**PART F. Provide a Panopto video recording that includes a demonstration of the functionality of the code used for the analysis and a summary of the tool(s) used.**

***The submission provides a Panopto video recording that accurately demonstrates the functionality of the code and summarizes the tools used.***

The session "D207 performance assessment" in folder "Data Cleaning NUM2 | D206 (student creators) [assignments]" is ready.  
  
You can view the session using the following link:  
<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0e5b0c12-e568-4121-be32-ae9300fe1d91>

**PART G. Reference the web sources used to acquire segments of third-party code to support the analysis.**

***The submission records all web sources used to acquire data or third-party code and all of the web sources are reliable.***

Deepanshu, B. (2020). *How to Import Data in Python.*

RSGB Business Consultant Pvt. Ltd. <https://www.listendata.com/2017/02/import-data-in-python.html#Import-CSV-files>

Pierre-Louis B. (2020). *Principle Components Analysis(PCA), Fundamentals, Benefits & Insights for Industry.*

Medium. <https://towardsdatascience.com/principal-components-analysis-pca-fundamentals-benefits-insights-for-industry-2f03ad18c4d7>

John S. (2018). *Data Cleaning with python and Pandas: Detecting Missing Values.*

Medium. <https://towardsdatascience.com/data-cleaning-with-python-and-pandas-detecting-missing-values-3e9c6ebcf78b>

Angelica Lo D. (2021). *How to detect outliers with Pytho pyod.*

Medium. <https://towardsdatascience.com/how-to-detect-outliers-with-python-pyod-aa7147359e4b>

Michael G. (2018). *Understanding Boxplots.*

Medium. <https://towardsdatascience.com/understanding-boxplots-5e2df7bcbd51>

**PART H. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.**

***The submission includes in-text citations for sources that are properly quoted, paraphrased, or summarized and a reference list that accurately identifies the author, date, title, and source location as available.***

**Reference**

Larose, C. D., & Larose, D. T. (2019). Data science using Python and R. ISBN-13: 978-1-119-52684-1.