IS THERE AN EFFECT OF BOTH INTERNET SERVICES AND ONLINE SECURITY ON MONTHLY CHARGE?

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

JUNE 29TH, 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, we aim to execute an investigation on a churn dataset to create a solution to a contemporary business inquiry such as – is there an effect of both internet services and online security on monthly charges?

A two-way anova test will be enforced in analyzing the relevant data sets because anova is a method to regulate on the occasion that the mean of groups appears to be distinctive in expected statistics, we operate sampling to figure out the characteristics of populations (Soner Yildirim, 2021).

***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 service positioned in revealing the effect of both online security and internet service on monthly charge. 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:

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.

**Age**: The specific age of the customer as reported during their subscription(18, 38, 46)

**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.***

|  |
| --- |
| import numpy as np  import pandas as pd  import scipy.stats as stat  import matplotlib.mlab as mlab  import matplotlib.pyplot as plt  import seaborn as sns  import statsmodels.api as sm  from statsmodels.formula.api import ols  from scipy import stats  %matplotlib inline  df = pd.read\_csv("churn\_clean.csv", usecols=['Varchar'] (input())  df  df.tail  df.info()  df.describe()  df.dtypes  df.shape  df.columns  df.isna()  df.isna().sum()  df.nunique()  df[)].hist()  boxplot=sns.boxplot(x='variable',data=df)  df['Varchar' (input () )].hist()  boxplot=sns.boxplot(x='Varchar' (input () ),data=df)  outlierFilter=df['Varchar' (input () )] (less than) int(input () )  df = df[outlierFilter]  boxplot=sns.boxplot(x='Varchar' (input () ),data=df)  pres\_duplicate=df.duplicated().sum()/df.shape[0]\*value  print('data rate of duplication {} %:'.format(pres\_duplicate))  df['Varchar' (input () )].hist()  boxplot=sns.boxplot(x='Varchar' (input () )].,data=df)  grouped'Varchar' (input () ) = df.groupby(by='Varchar' (input () ).size()  grouped'Varchar' (input () )  %matplotlib inline  grouped%matplotlib inline  grouped'Varchar' (input () ).plot.bar()  grouped'Varchar' (input () ) = df.groupby(by='Varchar' (input () ).size()  grouped'Varchar' (input () )  %matplotlib inline  grouped'Varchar' (input () ).plot.bar()  df.plot.scatter(x= 'Varchar' (input () ), y='Varchar' (input () )  df.plot.scatter(x= 'Varchar' (input () ), y='Varchar' (input () )  df.plot.scatter(x= 'Varchar' (input () ), y='Varchar' (input () )  df.plot.scatter(x= 'Varchar' (input () ), y='Varchar' (input () )  df.boxplot('Varchar' (input (), by='Varchar' (input ())  Ano2=ols('Varchar' (input ()~'Varchar' (input () )+'Varchar' (input () ),data=df).fit()  Two=sm.stats.anova\_lm(Ano2, typ=int (input ())  Two  Ano3=ols('Varchar' (input () )~ 'Varchar' (input () )+ 'Varchar' (input () )+ 'Varchar' (input () )\* 'Varchar' (input () ),data=df).fit()  Int=sm.stats.anova\_lm(Ano3, typ= int (input ())  Int |

***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.***

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| A screenshot of a computer  Description automatically generated with medium confidence  ***Table  Description automatically generated***  Text  Description automatically generated  Table  Description automatically generated  <class 'pandas.core.frame.DataFrame'>  RangeIndex: 10000 entries, 0 to 9999  Data columns (total 50 columns):  # Column Non-Null Count Dtype  --- ------ -------------- -----  0 CaseOrder 10000 non-null int64  1 Customer\_id 10000 non-null object  2 Interaction 10000 non-null object  3 UID 10000 non-null object  4 City 10000 non-null object  5 State 10000 non-null object  6 County 10000 non-null object  7 Zip 10000 non-null int64  8 Lat 10000 non-null float64  9 Lng 10000 non-null float64  10 Population 10000 non-null int64  11 Area 10000 non-null object  12 TimeZone 10000 non-null object  13 Job 10000 non-null object  14 Children 10000 non-null int64  15 Age 10000 non-null int64  16 Income 10000 non-null float64  17 Marital 10000 non-null object  18 Gender 10000 non-null object  19 Churn 10000 non-null object  20 Outage\_sec\_perweek 10000 non-null float64  21 Email 10000 non-null int64  22 Contacts 10000 non-null int64  23 Yearly\_equip\_failure 10000 non-null int64  24 Techie 10000 non-null object  25 Contract 10000 non-null object  26 Port\_modem 10000 non-null object  27 Tablet 10000 non-null object  28 InternetService 10000 non-null object  29 Phone 10000 non-null object  30 Multiple 10000 non-null object  31 OnlineSecurity 10000 non-null object  32 OnlineBackup 10000 non-null object  33 DeviceProtection 10000 non-null object  34 TechSupport 10000 non-null object  35 StreamingTV 10000 non-null object  36 StreamingMovies 10000 non-null object  37 PaperlessBilling 10000 non-null object  38 PaymentMethod 10000 non-null object  39 Tenure 10000 non-null float64  40 MonthlyCharge 10000 non-null float64  41 Bandwidth\_GB\_Year 10000 non-null float64  42 Item1 10000 non-null int64  43 Item2 10000 non-null int64  44 Item3 10000 non-null int64  45 Item4 10000 non-null int64  46 Item5 10000 non-null int64  47 Item6 10000 non-null int64  48 Item7 10000 non-null int64  49 Item8 10000 non-null int64  dtypes: float64(7), int64(16), object(27)  memory usage: 3.8+ MB  Table  Description automatically generated  Table  Description automatically generated  CaseOrder int64  Customer\_id object  Interaction object  UID object  City object  State object  County object  Zip int64  Lat float64  Lng float64  Population int64  Area object  TimeZone object  Job object  Children int64  Age int64  Income float64  Marital object  Gender object  Churn object  Outage\_sec\_perweek float64  Email int64  Contacts int64  Yearly\_equip\_failure int64  Techie object  Contract object  Port\_modem object  Tablet object  InternetService object  Phone object  Multiple object  OnlineSecurity object  OnlineBackup object  DeviceProtection object  TechSupport object  StreamingTV object  StreamingMovies object  PaperlessBilling object  PaymentMethod object  Tenure float64  MonthlyCharge float64  Bandwidth\_GB\_Year float64  Item1 int64  Item2 int64  Item3 int64  Item4 int64  Item5 int64  Item6 int64  Item7 int64  Item8 int64  dtype: object  (10000, 50)  Index(['CaseOrder', 'Customer\_id', 'Interaction', 'UID', 'City', 'State',  'County', 'Zip', 'Lat', 'Lng', 'Population', 'Area', 'TimeZone', 'Job',  'Children', 'Age', 'Income', 'Marital', 'Gender', 'Churn',  'Outage\_sec\_perweek', 'Email', 'Contacts', 'Yearly\_equip\_failure',  'Techie', 'Contract', 'Port\_modem', 'Tablet', 'InternetService',  'Phone', 'Multiple', 'OnlineSecurity', 'OnlineBackup',  'DeviceProtection', 'TechSupport', 'StreamingTV', 'StreamingMovies',  'PaperlessBilling', 'PaymentMethod', 'Tenure', 'MonthlyCharge',  'Bandwidth\_GB\_Year', 'Item1', 'Item2', 'Item3', 'Item4', 'Item5',  'Item6', 'Item7', 'Item8'],  dtype='object')  Table  Description automatically generated  Table  Description automatically generated  CaseOrder 0  Customer\_id 0  Interaction 0  UID 0  City 0  State 0  County 0  Zip 0  Lat 0  Lng 0  Population 0  Area 0  TimeZone 0  Job 0  Children 0  Age 0  Income 0  Marital 0  Gender 0  Churn 0  Outage\_sec\_perweek 0  Email 0  Contacts 0  Yearly\_equip\_failure 0  Techie 0  Contract 0  Port\_modem 0  Tablet 0  InternetService 0  Phone 0  Multiple 0  OnlineSecurity 0  OnlineBackup 0  DeviceProtection 0  TechSupport 0  StreamingTV 0  StreamingMovies 0  PaperlessBilling 0  PaymentMethod 0  Tenure 0  MonthlyCharge 0  Bandwidth\_GB\_Year 0  Item1 0  Item2 0  Item3 0  Item4 0  Item5 0  Item6 0  Item7 0  Item8 0  dtype: int64  CaseOrder 10000  Customer\_id 10000  Interaction 10000  UID 10000  City 6058  State 52  County 1620  Zip 8583  Lat 8563  Lng 8630  Population 5933  Area 3  TimeZone 25  Job 639  Children 11  Age 72  Income 9993  Marital 5  Gender 3  Churn 2  Outage\_sec\_perweek 9986  Email 23  Contacts 8  Yearly\_equip\_failure 6  Techie 2  Contract 3  Port\_modem 2  Tablet 2  InternetService 3  Phone 2  Multiple 2  OnlineSecurity 2  OnlineBackup 2  DeviceProtection 2  TechSupport 2  StreamingTV 2  StreamingMovies 2  PaperlessBilling 2  PaymentMethod 4  Tenure 9996  MonthlyCharge 750  Bandwidth\_GB\_Year 10000  Item1 7  Item2 7  Item3 8  Item4 7  Item5 7  Item6 8  Item7 7  Item8 8  dtype: int64  <AxesSubplot:> MonthlyCharge  <AxesSubplot:>      <AxesSubplot:> Age      InternetService  DSL 3293  Fiber Optic 4204  None 2033  dtype: int64  <AxesSubplot:xlabel='InternetService'>    OnlineSecurity  No 6104  Yes 3426  dtype: int64  <AxesSubplot:xlabel='OnlineSecurity'>    <AxesSubplot:xlabel='MonthlyCharge', ylabel='InternetService'>    <AxesSubplot:xlabel='Age', ylabel='InternetService'>    <AxesSubplot:xlabel='MonthlyCharge', ylabel='OnlineSecurity'>    <AxesSubplot:xlabel='Age', ylabel='OnlineSecurity'>      <AxesSubplot:title={'center':'MonthlyCharge'}, xlabel='InternetService'>     |  | **sum\_sq** | **df** | **F** | **PR(>F)** | | --- | --- | --- | --- | --- | | **InternetService** | 1.676708e+06 | 2.0 | 503.306181 | 1.696981e-208 | | **OnlineSecurity** | 3.104131e+04 | 1.0 | 18.635657 | 1.598361e-05 | | **Residual** | 1.583909e+07 | 9509.0 | NaN | NaN |  |  | **sum\_sq** | **df** | **F** | **PR(>F)** | | --- | --- | --- | --- | --- | | **InternetService** | 1.676708e+06 | 2.0 | 503.364597 | 1.617571e-208 | | **OnlineSecurity** | 3.104131e+04 | 1.0 | 18.637820 | 1.596556e-05 | | **InternetService:OnlineSecurity** | 5.169150e+03 | 2.0 | 1.551830 | 2.119135e-01 | | **Residual** | 1.583392e+07 | 9507.0 | NaN | NaN | |

The hypothesis:

Null hypothesis H0- There is no effect on both online security and Internet service on monthly charge.

Alternative hypothesis H1: There is an effect on both online security and Internet service on monthly charge

The f-statistic corresponding to Online security is 18.63 which is not significant at p<1.598. The F-statistics for Internet service is 503.3 which is significant at p<1.696. Therefore, the interaction between online

security and internet service are not significant as well.

However, (H1) is not supported, while the initial online security doesn’t have a significant effect on the

monthly charge alongside the internet service which appears not to have a considerable outcome.

In Addition, online security has no significant effect on customers monthly charge, F (1.0, 1.583) = 18.63, P =1.596.The internet service also doesn’t have a 100% effect on monthly charge, F (2., 1.583) = 503.3, P =1.61.

The interaction of internet service and online security does not affect the monthly charge of customers, F (2, 1.583)

= 1.551, P =2.11 respectively.

***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 reason why we chose anova:

* Statistical tests such as ANOVA support us rationalize if outcome examined are associable to populations.
* ANOVA could be spread-out to three or further association.
* ANOVA could additionally be passed down in the feature election technique of machine learning. the features could possibly be correlated by carrying-out an anova test alongside identical ones could be dropped in distinction to the feature set.

**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 – Age and MonthlyCharge**
* **Categorical variables - Internet service and online security**

**2 Continuous variables – Age and MonthlyCharge**

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

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

|  |
| --- |
| %matplotlib inline  groupedInternetService.plot.bar()  <AxesSubplot:xlabel='InternetService'>    %matplotlib inline  groupedOnlineSecurity.plot.bar()  <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 – Age and MonthlyCharge**
* **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.**

|  |
| --- |
| sns.scatterplot(x='InternetService', y='OnlineSecurity', data=df)  plt.show() |

2 Continuous variables – Age and MonthlyCharge

|  |
| --- |
| sns.scatterplot(x='Age', y='MonthlyCharge', data=df)  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 two way anova testing method below is the result.

The hypothesis:

Null hypothesis H0- There is no effect on both online security and Internet service on monthly charge.

Alternative hypothesis H1: There is an effect on both online security and Internet service on monthly charge

The f-statistic corresponding to Online security is 18.63 which is not significant at p<1.598. The F-statistics for Internet service is 503.3 which is significant at p<1.696. Therefore, the interaction between online

security and internet service are not significant as well.

However, (H1) is not supported, while the initial online security doesn’t have a significant effect on the

monthly charge alongside the internet service which appears not to have a substantial effect.

In Addition, online security has no significant effect on customers monthly charge, F (1.0, 1.583) = 18.63, P =1.596.The internet service also doesn’t have a 100% effect on monthly charge, F (2., 1.583) = 503.3, P =1.61.

The interaction of internet service and online security does not affect the monthly charge of customers, F (2, 1.583)

= 1.551, P =2.11 respectively.

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.**

* One constraint of anova is that it could exclusively be practiced to analyze the means of two or higher clusters.
* It is not analytically possible to be applied in creating a contrast of the means of greater than two clusters.
* A further restriction of anova is that it appears not to be constantly applicable for an entire kind of data.

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 subscriptions and online security preferences matched together in reference to the effects of both online security and internet service on the monthly charge of the customers.

**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.***

You can view the session using the following link:  
<https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=f0bb0eef-eb23-4544-9426-aec100cfa0ea>

**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.***

Yildirim, S. (2021). *What is ANOVA and When Do We Use It?*

Medium [***https://towardsdatascience.com/what-is-anova-and-when-do-we-use-it-e0dcd8759c4e***](https://towardsdatascience.com/what-is-anova-and-when-do-we-use-it-e0dcd8759c4e)

Date, S. (2021). *The Chi-squared Goodness of Fit Test for Regression Analysis.*

Medium. <https://medium.com/towards-data-science/the-chi-squared-goodness-of-fit-test-for-regression-analysis-b2d2e4b487ab>

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.