D207- Exploratory Data Analysis Performance Assessment

Western Governs University

Table of Contents

Part I:

A1. Question for Analysis

Do customers experiencing the most outages with their service and equipment failure increase the cancellation of service? Thus my research question will be: does outages per week and yearly equipment failure impact churn?

A2. Benefits from Analysis

Customers can choose from multiple telecommunication services as their provider. Customer churn is the percentage of customers who discontinued their services from a provider within a specific time frame. The churn rate in the data set is documented by customers who discontinued their services within the last month. By examining the reasons behind customer churn stakeholders can identify patterns in customer behaviors and preferences for increased customer retention. This can lead to decreased churn rates and increased business profits.

A3. Data Identification

The data was previously cleaned using python by detecting duplicate data, missing values, outliers, and any other data quality issues in the churn data set. To start the data cleaning process he data types included in the churn data were determined. The data type of each variable is needed information due to certain functions working only with specific functions. This includes the column names and the number of non-null values for each column. Once datatypes are known the data could be cleaned. Cleaning data included detecting duplicates, and identifying missing values and outliers. There were no duplicates in the data set. The next step included determining if there were missing values. Missing values are usually represented in the form of nan, null, or none in the dataset. The following columns included missing values: children, age, income, techie, phone, tech support, tenure, bandwidth_GB_year. Theses variavles included quantitative data. The columns techie, phone, and techsupport also had missing data. These columns have qualitative data consisting of YES/NO values. Ordinal encoding was used to re-express values as numeric values. Univariate imputation was then used to treat missing values for all quantitative data. The mean and median were calculated to replace missing values.

All the quantitative variables from the churn data were plotted on boxplots to visualize outliers. The exact number of outliers was determined using the z-values since the boxplot could not provide an exact number. Outliers were then treated using the retention method. Outliers that were more than 3 standard deviations above the mean were removed. Lastly, PCA was performed with the numerical variables from the data set for increased data compression and visualization.

By cleaning the data set the variable churn was determined to be the dependent variable and is binary categorical. An independent variable is a variable that is being manipulated or controlled and a dependent variable is a variable that is being studied for changes based on the independent variable (Bruce et al., 2020) . Binary categorical means there are two categorical variables and only two possible values: "yes" and "no". The churn variable indicates if customers dropped their services in the telecommunication company in the last month by answering "yes" and "no". There are 49 other variables in the churn data set and these are the independent variables. The independent variables of focus for the analysis will be outage_sec_perweek and yearly_equip_failure to answer the research question. These are both quantitative (continuous) variables while the churn variable is a qualitative (categorical) variable. Thus the ANOVA test was used to find the possibility of a positive relationship between the independent and dependent variables. ANOVA is a technique used to compare the mean values between three or more groups to determine if there is a statical significance (Bruce et al., 2020).

Part II:

B1. Code

Kolmogorov-Smirnov Test Code:

```
from scipy.stats import kstest, norm
import numpy as np
np.random.seed(123)
sample = np.random.normal(loc=0, scale=1, size=100)
ks_stat, p_value = kstest(sample, norm.cdf)
print("Kolmogorov-Smirnov test:")
print("KS statistic:", ks_stat)
print("p-value:", p_value)
```

Linear Regression Analysis Code:

```
from sklearn.linear_model import LinearRegression
model = LinearRegression()
model.fit(X_train, y_train)
score = model.score(X_test, y_test)
print("Model score:", score)
new_data = {'Outage_sec_perweek': [10.5, 20.1, 5.3],
        'Yearly_equip_failure': [0, 2, 1]}
X_new_data = pd.DataFrame(new_data)
y_pred = model.predict(X_new_data)
print("Predictions:", y_pred)
```

ANOVA with Linear Regression Test Code:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols
model = LinearRegression()
model.fit(X, y)
formula = 'Churn ~ Outage_sec_perweek + Yearly_equip_failure'
anova_model = ols(formula, data=df).fit()
print(anova_model.summary())
```

ANOVA Test:

```
import statsmodels.api as sm
from statsmodels.formula.api import ols

model = ols('Churn ~ Outage_sec_perweek + Yearly_equip_failure', data=df).fit()
anova_table = sm.stats.anova_lm(model, typ=2)
print(anova_table)
```

B2. Output:

ANOVA with linear regression:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                  Churn   R-squared:                       0.000
Model:                            OLS   Adj. R-squared:                 -0.000
Method:                 Least Squares   F-statistic:                    0.6117
Date:                Wed, 15 Feb 2023   Prob (F-statistic):              0.542
Time:                        23:00:38   Log-Likelihood:                -6009.2
No. Observations:               10000   AIC:                         1.202e+04
Df Residuals:                    9997   BIC:                         1.205e+04
Df Model:                           2
Covariance Type:            nonrobust
==============================================================================
                             coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept                  0.2814      0.017     16.620      0.000       0.248       0.315
Outage_sec_perweek        -0.0014      0.002     -0.899      0.369      -0.004       0.002
Yearly_equip_failure      -0.0060      0.009     -0.633      0.527      -0.024       0.012
==============================================================================
Omnibus:                     2459.203   Durbin-Watson:                   1.521
Prob(Omnibus):                  0.000   Jarque-Bera (JB):             2201.754
Skew:                           1.065   Prob(JB):                         0.00
Kurtosis:                       2.134   Cond. No.                         41.8
==============================================================================

Notes:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

ANOVA Test:

	sum_sq	df	F	PR(>F)
Outage_sec_perweek	0.157460	1.0	0.808274	0.368653
Yearly_equip_failure	0.078115	1.0	0.400980	0.526598
Residual	1947.511655	9997.0	NaN	NaN

B3. Justification

ANOVA was chosen to analyze this data set because this technique is used to compare the means between three or more groups to determine statistical significance between the groups. The research questions includes three variables. In an ANOVA test, the dependent variable is quantitative and the independent variable(s) are categorial. The independent variables of focus for the analysis was outage_sec_perweek and yearly_equip_failure and the dependent variable was churn. ANOVA assumes that the data is normally disturbed and the variances are equal

across all groups (LabXchange, n.d.). A Kolmogorov-Smirnov test was completed to test the data set for normality prior to running the linear regression and ANOVA test.

Part III

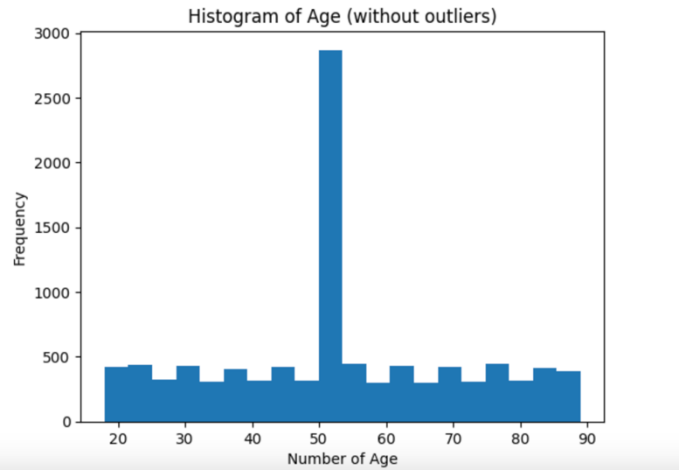
C. Univariate Statistics

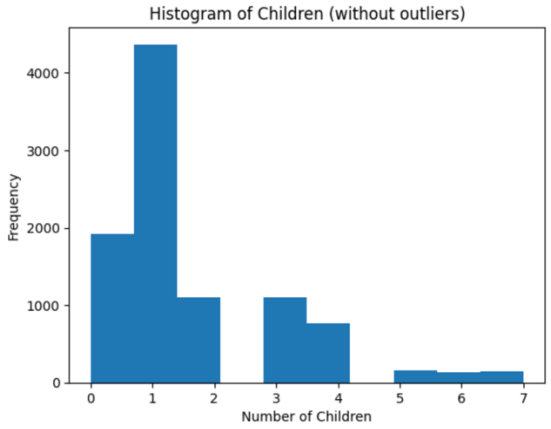
Below is the distribution of two continuous variables and two categorical variables using univariate statistics. Univariate statistics is the statical analysis of a single variable at one time (Bruce et al., 2020). The two continuous variables chosen were 'Income' and 'Age'. A histogram was used to visualize the distribution of these two variables. The distribution of the 'Income' variable indicated a right-skewed distribution (LabXchange, n.d.). The 'age' variable indicated a uniform distribution.

When examining the distribution of categorical variables using univariate statistics a bar chart was used. The categorical variables chosen were 'Gender' and 'phone'. The 'Gender' variable had a relatively uniform distribution with "female" and "male" representation almost equal (LabXchange, n.d.). When factoring in the "prefer not to answer" category the distribution becomes skewed toward the male and female categories. The 'phone' variable was skewed with most of the responses being "yes".

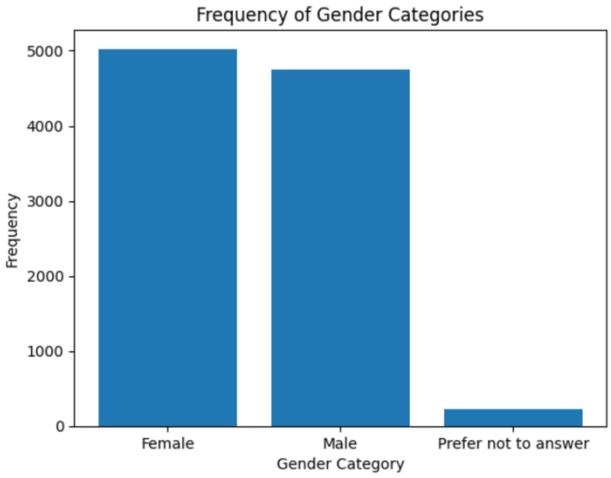
C1. Visual of Findings

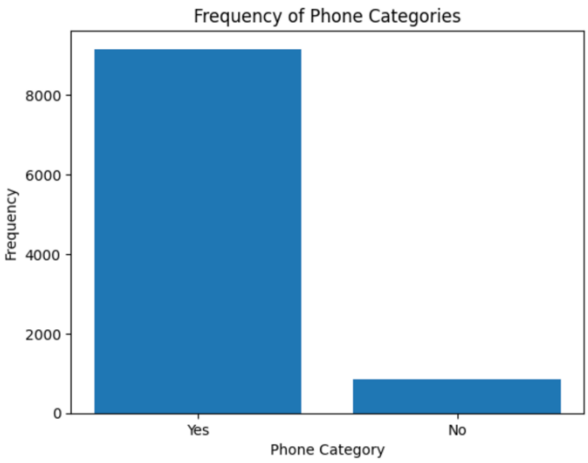
Continuous variables:





Categorical Variables:





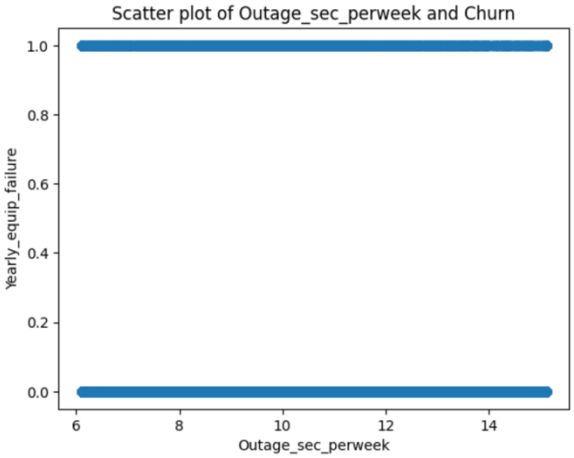
Part IV:

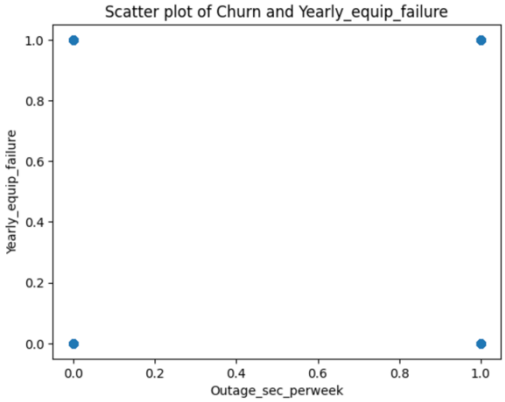
D. Bivariate Statistics

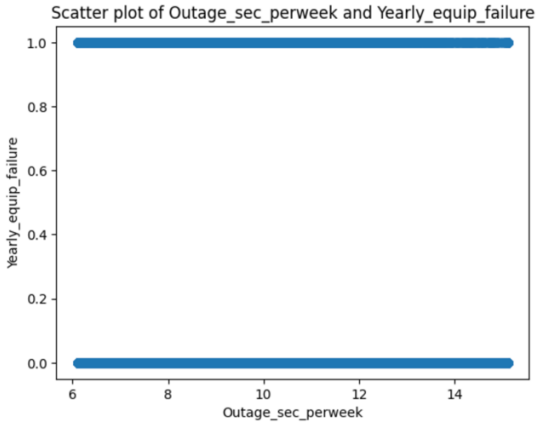
Bivariate statistical analysis refers to the statistical analysis of two variables at once (Bruce et al., 2020). This bivariate analysis includes the distribution of two continuous variables and two categorical variables. The two continuous variables chosen were ‘outage_sec_perweek’ and ‘yearly_equip_failure’. A scatterplot was also created for the ‘churn’ variable to examine the relationship with each of the independent variables ‘outage_sec_perweek’ and ‘yearly_equip_failure’. The categorial variables chosen were “churn” and analyzed against ‘outage_sec_perweek’ and ‘yearly_equip_failure’. The bar chart from an ANOVA analysis below shows a skewed distribution (LabXchange, n.d.).

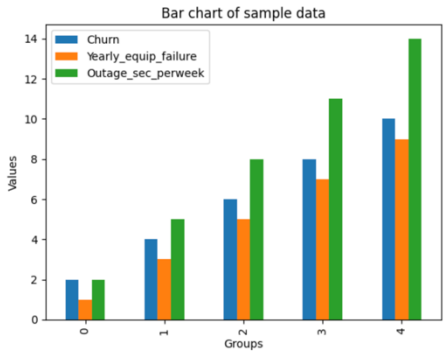
.

D1. Visual of Findings









Part V:

E1. Results of Analysis

The null hypothesis was not rejected. The variables ‘outage_sec_perweek’ and ‘yearly_equip_failure’ showed no significant relationship to the ‘churn’ variable. The alternative hypothesis would be that these two independent variables are dependent on the dependent variable. The ANOVA test provided a p-value greater than 0.05 indicating no significant correlation. The p-value is "the frequency with which the chance model produces a result more extreme than the observed result "(Bruce, Bruce, et al., 2020).

By examining the bivariate analysis we can confirm the null hypothesis cannot be rejected. The scatterplot and box plots visually confirmed there was no significance between the independent variables and the dependent variable.

E2. Limitation of Analysis

The null hypothesis could not be rejected so there was no evidence to support the claim of ‘outage_sec_perweek’ and ‘yearly_equip_failure’ having a direct correlation/impact on the dependent variable ‘churn’. This result could be due to various reasons. Some limitations of my analysis could include sample size. The sample size could have been too small to detect a difference between the independent and dependent variables. Confounding variables could also affect the outcome of this analysis. Variables not studied in this analysis could have led to the differences in customer satisfaction. Another limitation in the analysis could have been if there was too much variation in the data to show a correlation.

E3. Recommended Course of Action

Since the null hypothesis could not be rejected so further analysis would need to be done. One course of action would be to analyze different independent variables again the dependent variable ‘churn’. This analysis did not include any independent variables that utilized customer reported data. Customer reported variables could show a correlation to the dependent variable. Another course of action would be to increase the sample size over a longer period of time. Additional analysis could also be done. A different type of analysis of the data or a different statical method could be completed to measure the variables.

Part VI

F. Panopto Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=0b6d0446-10ee-4514-93f7-afac0102268d

Part VII

G. Sources for Third-Party Code:

Matplotlib Bar chart. (2016, July 28). Pythonspot. https://pythonspot.com/matplotlib-bar-chart/

(Matplotlib Bar Chart, 2016)

H. References:

Bruce, P., Bruce, A., & Gedeck, P. (2020). *Practical Statistics for Data Scientists: 50+ Essential Concepts Using R and Python*. O'Reilly Media.

LabXchange.(n.d.). https://www.labxchange.org/library/items/lb:LabXchange:10d3270e:html:1