# Exploratory Data Analysis D207

January 4, 2022

## 1 Part A: Real-World Organizational Situtation or Issue

#### 1.0.1 1: How does the realibily of the customer's equipment and the timeliness of fixing and replacing equipment affect the Churn Rate of customers?

#### 1.0.2 2: Stakeholders in the organization can benefit from this by knowing if allocating resources to improving the timeliness of fixing and replacing of equipment can reduce the Churn Rate of their customers.

#### 1.0.3 3: The relevant data needed to answer this question is as follows.

- Did the customer Churn?
- Were the responses to fixes timely?
- Were the responses to replacements timely?
- What was the average number of seconds per week of system outages in the customer's neighborhood?
- What was the number of times customer's equipment failed and had to be reset/replaced in the past year?

## 2 Part B: Data Analysis using a chi-square

#### 2.0.1 1. Code to run the analysis

[1]:
```
# Imports

import pandas as pd
import matplotlib.pyplot as plt
%matplotlib inline
import seaborn as sns

from scipy import stats
from scipy.stats import chi2
from scipy.stats import chi2_contingency


# Raw Data to DataFrame
df = pd.read_csv('churn_clean.csv', encoding='utf-8', index_col=0)
```

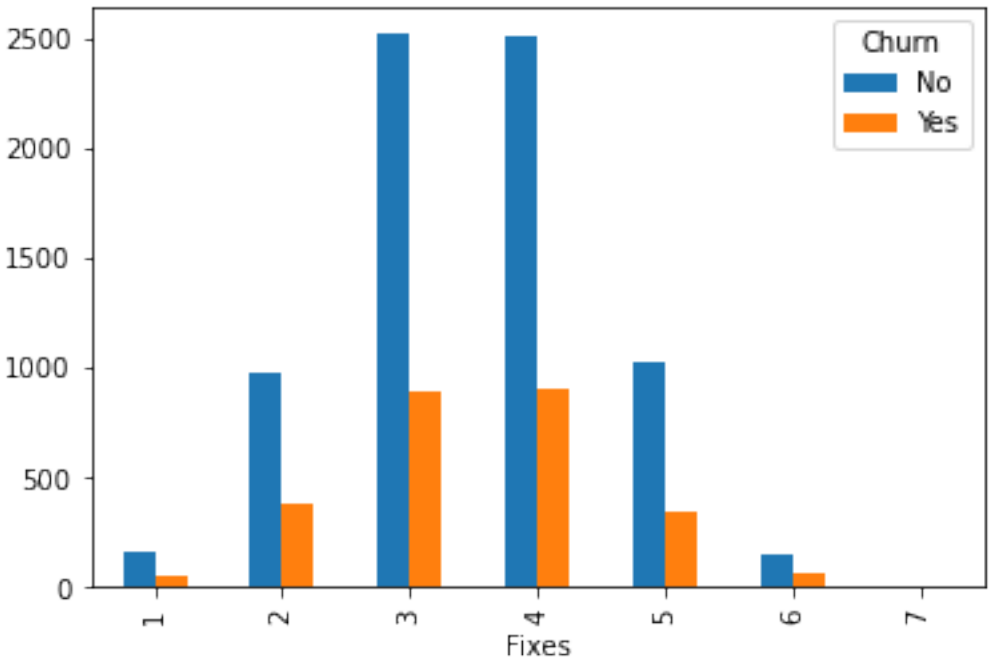
[2]:
```
# Rename survey columns for better readability
df.rename(columns = {
                     'Item2':'Fixes',
                      'Item3':'Replacements', },
          inplace=True)
```

[3]:
```
ctab_fixes = pd.crosstab(df['Fixes'], df['Churn'])
ctab_fixes
```

[3]:
```
Churn     No  Yes
Fixes
1        160   57
2        973  387
3       2519  896
4       2507  905
5       1025  343
6        155   60
7         11    2
```

[4]:
```
ctab_fixes.plot(kind='bar', stacked=False)
```

[4]:
```
<AxesSubplot:xlabel='Fixes'>
```

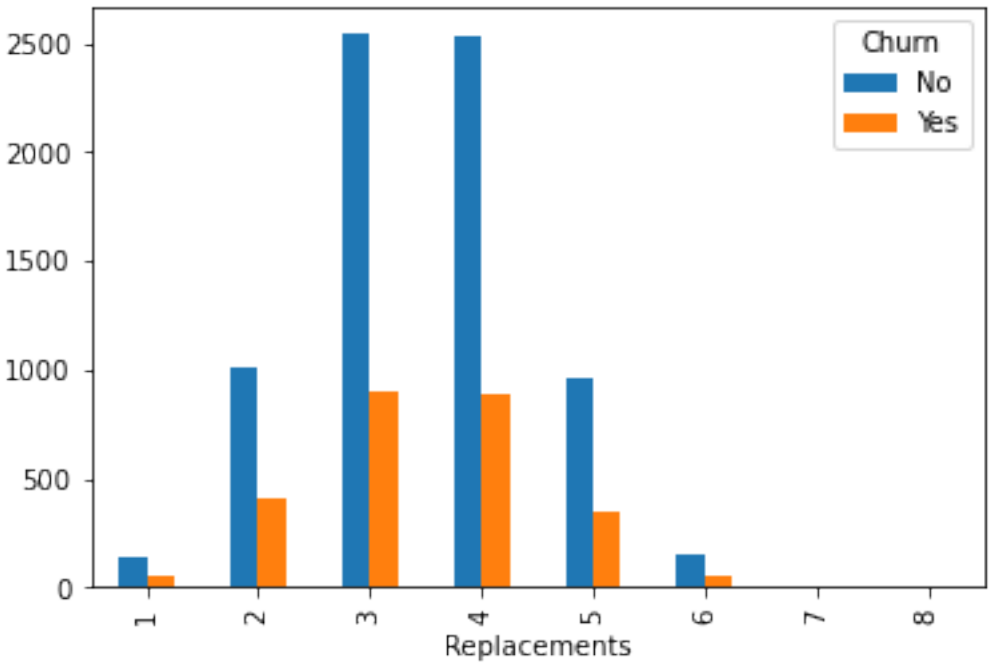



[5]:
```
ctab_replacements = pd.crosstab(df['Replacements'], df['Churn'])
ctab_replacements
```

[5]:
| Churn | No | Yes |
|---|---|---|
| Replacements | | |
| 1 | 146 | 56 |
| 2 | 1017 | 407 |
| 3 | 2540 | 895 |
| 4 | 2527 | 883 |
| 5 | 960 | 353 |
| 6 | 149 | 54 |
| 7 | 10 | 2 |
| 8 | 1 | 0 |

[6]:
```
ctab_replacements.plot(kind='bar', stacked=False)
```

[6]: <AxesSubplot:xlabel='Replacements'>




### 2.0.2 2. Output of calculations

Here we can do chi-square tests for independence. This will be tell us if the yes/no variable Churn is independent from our Timely Fixes/Replacement variables. (Brownlee, J. 2019)

[7]:
```
# Chi-square test of independence for fixes
stat, p, dof, expected = chi2_contingency(ctab_fixes)
print('Churn/Fixes p value = ' + str(p))
```

```
Churn/Fixes p value = 0.5093789499498207
```

[8]:
```
# Chi-square test of independence for replacements
stat, p, dof, expected = chi2_contingency(ctab_replacements)
print('Churn/Replacements p value = ' + str(p))
```

```
Churn/Replacements p value = 0.6148391285975547
```

#### 2.0.3 3. Justification

I used the chi-squared technique to test the dependancy between the categorical variable of whether or not a customer Churned, to the categorical variables Timely Fixes/Replacements. By using chi-squared technique I was able to determine how significant the variables were to eachother and if there was a significate correlation between them. (Brownlee, J. 2019)

## 3 Part C: Identify the distribution of two continuous variables and two categorical variables using univariate statistics

#### 3.0.1 Continuous variables:

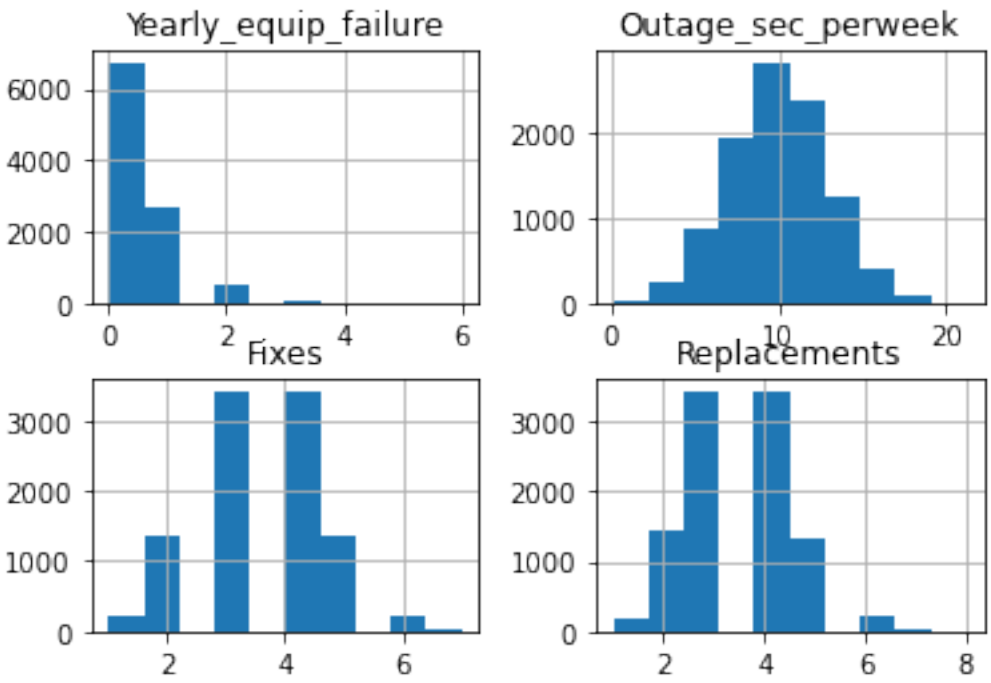
- Outage seconds per week
- Yearly equipment failure

#### 3.0.2 Categorical variables:

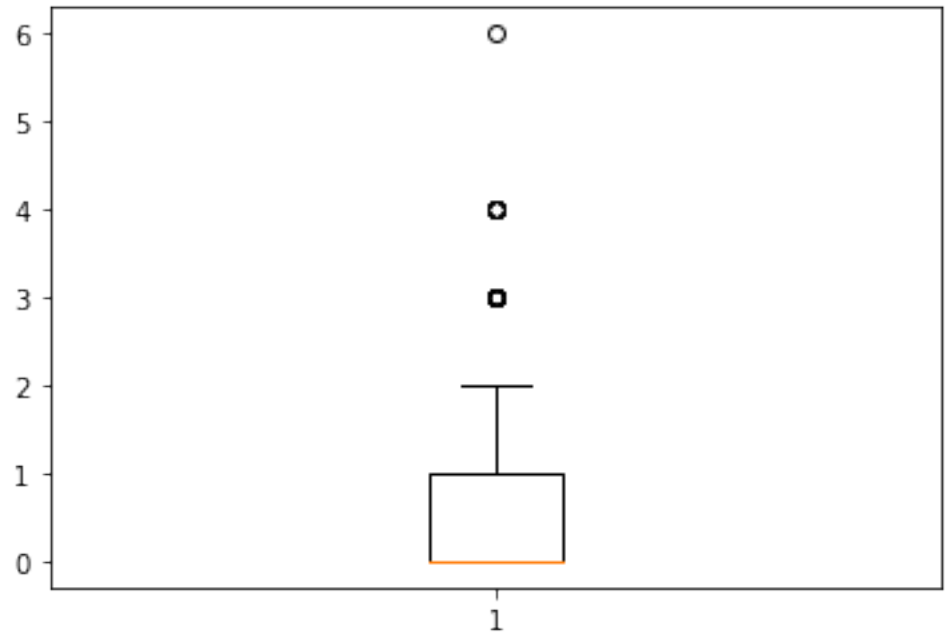
- Timely Fixes
- Timely Replacements

#### 3.0.3 1. Visual representations

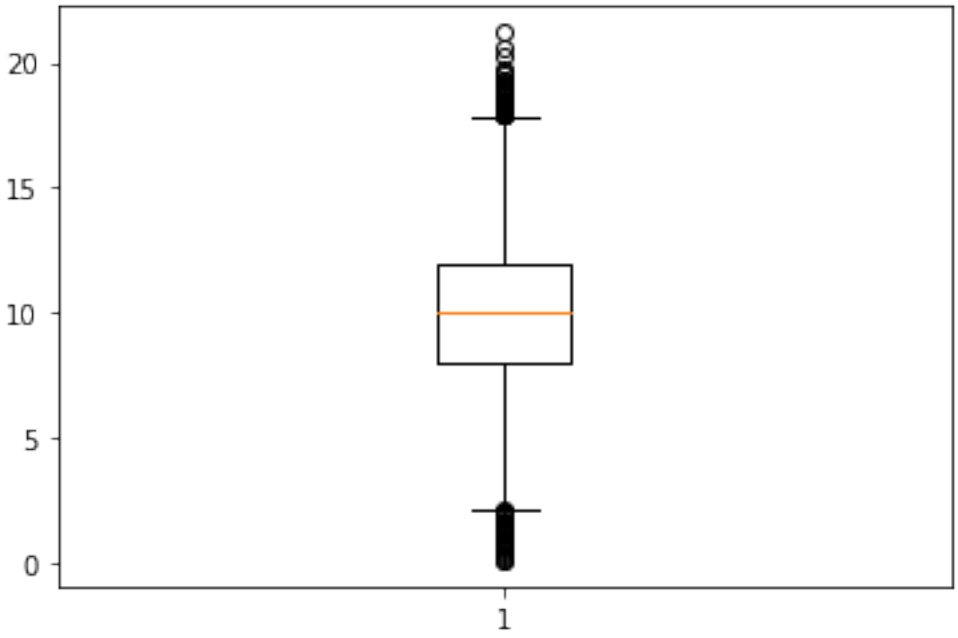
[9]:
```
df[['Yearly_equip_failure', 'Outage_sec_perweek', 'Fixes', 'Replacements']].
↪hist()
plt.show()
```



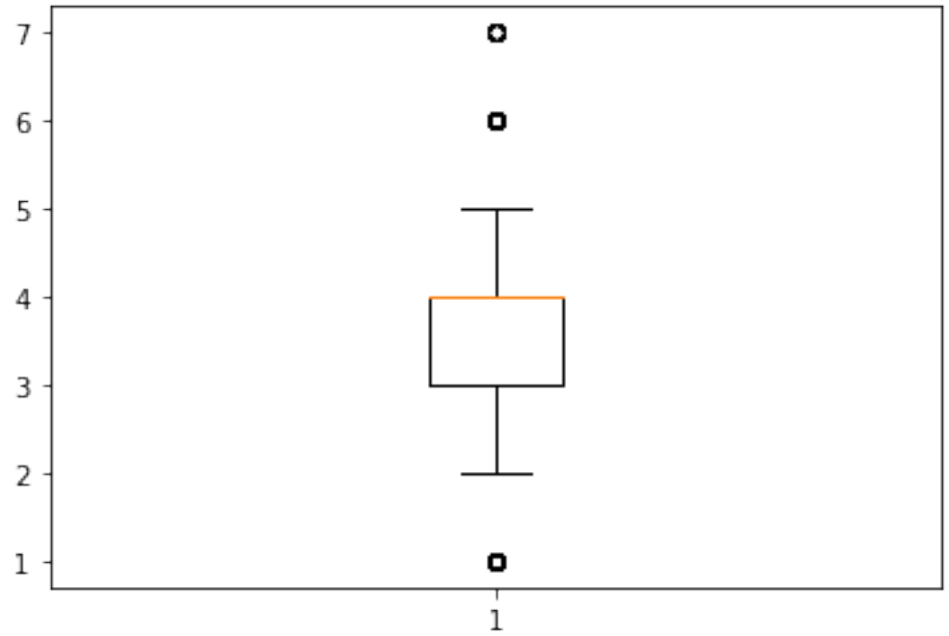

[10]:
```
plt.boxplot(df['Yearly_equip_failure'])
plt.show()
```



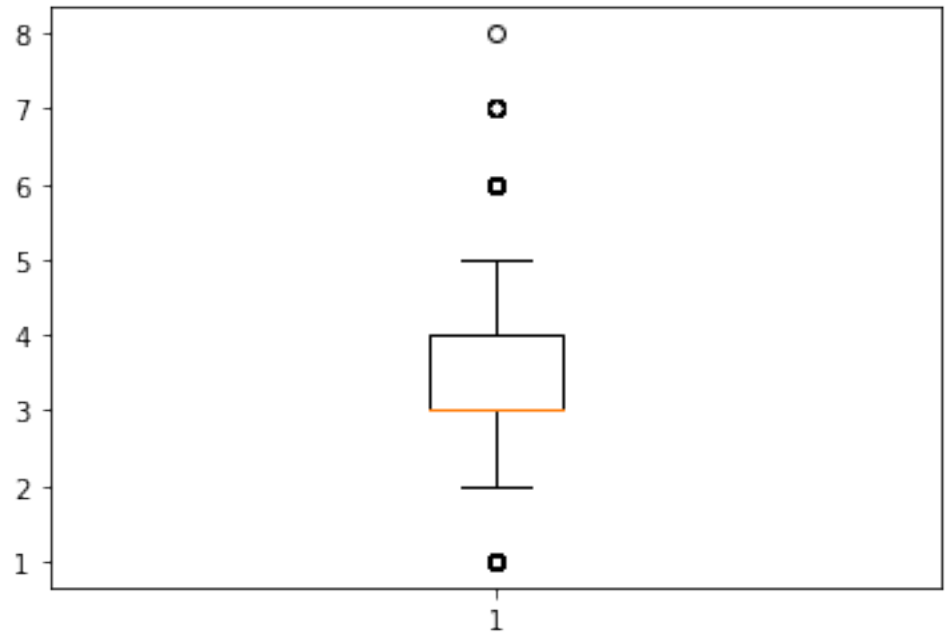
[11]:
```
plt.boxplot(df['Outage_sec_perweek'])
plt.show()
```




[12]:
```
plt.boxplot(df['Fixes'])
plt.show()
```



[13]:
```
plt.boxplot(df['Replacements'])
plt.show()
```




# 4 Part D: Identify the distribution of two continuous variables and two categorical variables using bivariate statistics

### 4.0.1 Continuous variables:

- Outage seconds per week
- Yearly equipment failure

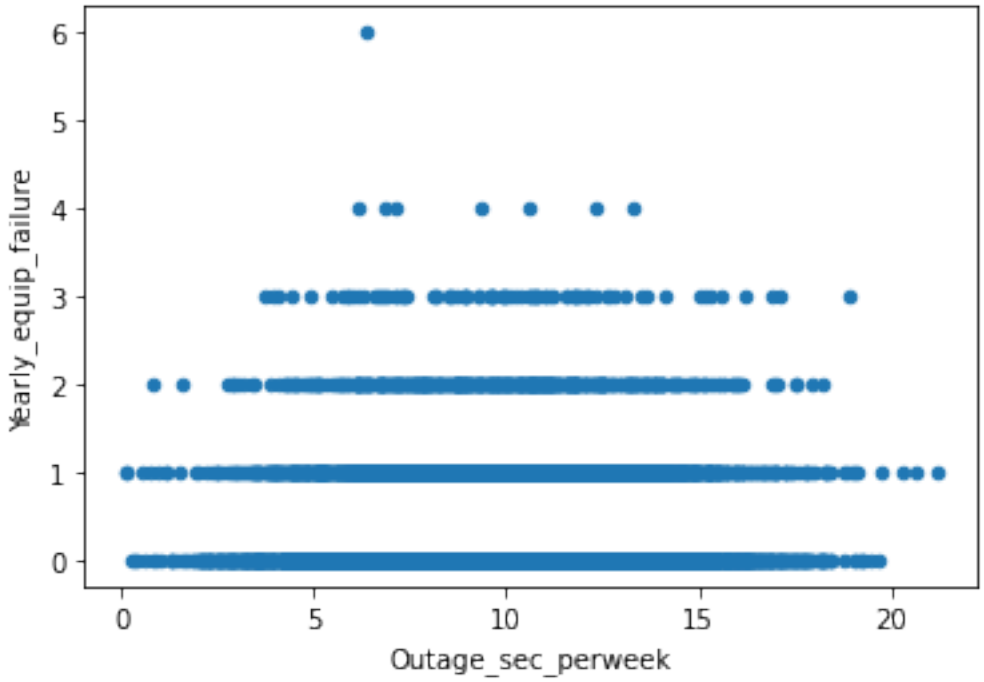
### 4.0.2 Categorical variables:

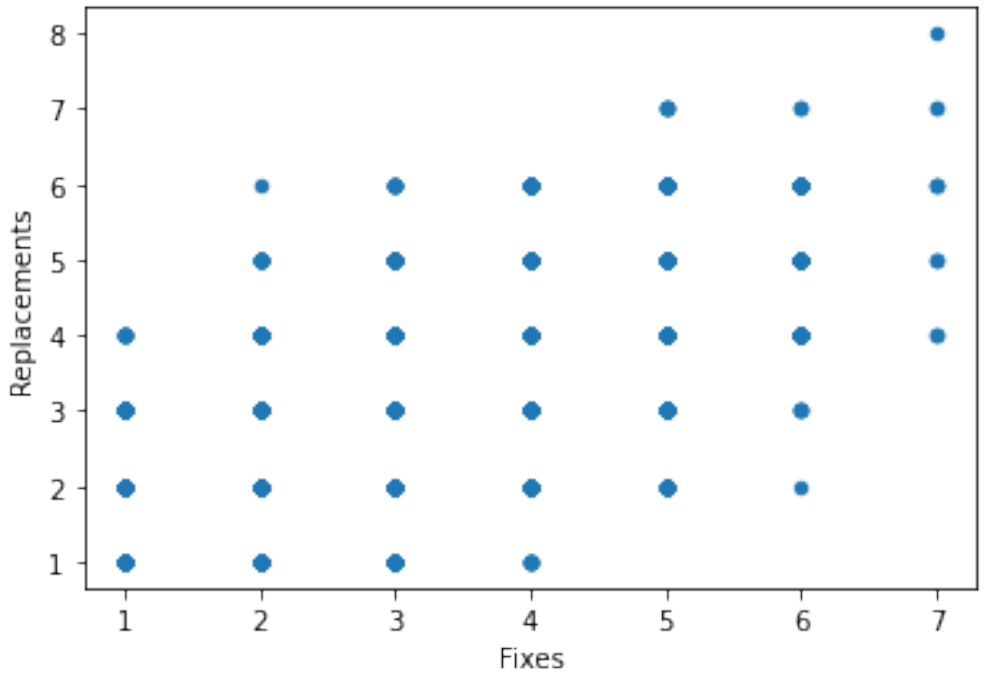
- Timely Fixes
- Timely Replacements

[14]:
```
bivariate_variables = df[['Yearly_equip_failure', 'Outage_sec_perweek',␣
↪'Fixes', 'Replacements']]
```

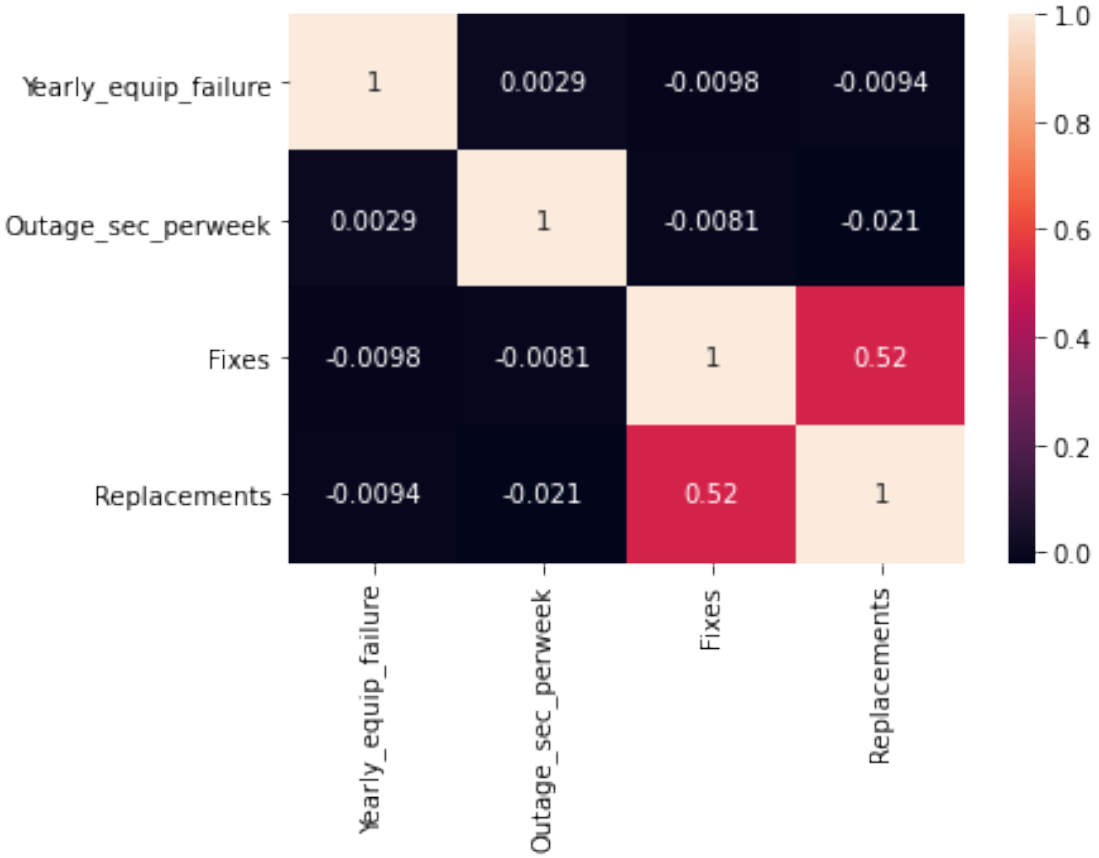
[15]:
```
## Scatter plots generated according to Bilogur, A. (2018)

bivariate_variables.plot.scatter(x='Outage_sec_perweek',␣
↪y='Yearly_equip_failure')
plt.show()
```

```
bivariate_variables.plot.scatter(x='Fixes',  y='Replacements')
plt.show()
```



[16]:
```
sns.heatmap(bivariate_variables.corr(), annot=True)
plt.show()
```




# 5 Part E: Summarize the implications

### 5.0.1 1. Results

The p values found from both chi-square testing: - Churn/Fixes p value = 0.5093789499498207 - Churn/Replacements p value = 0.6148391285975547 Both results are high enough to where we cannot reject a null hypothesis of the variables being independent at a standard significance alpha of .05. Given the data available, it's not clear if there's a statistically significant relationship between the timeliness of the customer's equipment being fixed or replaced with if they're more likely to churn or not. (Brownlee, J. 2019)

### 5.0.2 2. Limitations

Without more information surrounding the survey questions it's difficult to determine how dependable they are for further analysis or for analysis with other variables in this dataset.

#### 5.0.3 3. Recommended course of action

While fixing and replacing equipment in a timely manner would seem important to customer satisfaction, the data analysis on these variables show that there isn't statistical significance regarding these variables and how often customers have churned. The recommended course of action is continue to provide services in a timely manner to the customer, and to continue to analyze different variables for other trends that are more clearly significant statistically to act on.

## 6 Part F: Video

https://wgu.hosted.panopto.com/Panopto/Pages/Viewer.aspx?id=e41699f3-55ef-4471-a4ec-ae10017c8d6d

## 7 Part G: Code Sources

Bilogur, A. (2018, September 19). *Bivariate plotting with Pandas.* Kaggle. Retrieved January 4, 2022, from https://www.kaggle.com/residentmario/bivariate-plotting-with-pandas

**Part H: Sources** Brownlee, J. (2019, October 30). *A gentle introduction to the chi-squared test for machine learning.* Machine Learning Mastery. Retrieved January 4, 2022, from https://machinelearningmastery.com/chi-squared-test-for-machine-learning/

[ ]: