Part 1: Research Question

1. Which services predict the length of time (Tenure) a customer remains with the company?

2. The objective of the analysis is to determine which services are being used by customers with higher Tenure. Knowledge of this will allow the company to market high-tenure services to current customers not using them. Also, the company can use this information to attract new customers with offers that include these services.

Part 2: Method Justification

1. When using multiple regression, four assumptions are made. The first is linearity; the relationship between the outcome variable and the explanatory variables must be linear. Next, there should be homoscedasticity. The variance of the residuals is constant for all variables. Third, residuals are normally distributed. Finally, there should be no multicollinearity between the explanatory variables. (Using Statistical Regression Methods in Education Research, 2011)

2. Python was the language of choice for many reasons. It offers the versatility to do simple calculations as well as complex ones. It works across all platforms – Macs, Windows, Linux. Python runs faster than R. Most importantly, Python has an extensive library of packages like Numpy, Scikit-learn, and Mathplotlib. These packages allow Python to handle data analysis and machine learning efficiently. (Massaron & Boschetti, 2016)

3. Multiple Regression is the appropriate technique because the data set contains both continuous and categorical variables. It shows the relationship between the explanatory and outcome variables and puts it into a mathematical formula. The most practical reason for the telecommunication company is the ability to use multiple regression to make predictions. The company can use this information to decide what services and products are most important to retaining customers for a longer timeframe.

Part 3: Data Preparation

1. There are several steps to preparing the data for analysis. First, remove unnecessary columns from the data set. The raw data contains 50 columns, some that are not relevant to answering the research question. Next, look for missing values, duplicate values, or outliers. Then, enumerate all categorical data using dummies. This step is necessary because multiple linear regression does not allow for categorical variables. Finally, both the prince package and the statsmodel package require numeric values to be float types; therefore, all variables are converted to float64.

2. The target variable for this question is Tenure. The average is about 35 months, with a minimum of one month and a maximum of just under 72 months. The continuous numeric

independent variables include: Income, Outage_sec_perweek, MonthlyCharge, Bandwidth_GB_Year. Income has a mean of about $39806 annually with a minimum of $348 and a maximum of $258900. Outage_sec_perweek is the average number of seconds per week that the customer's neighborhood experiences a service outage. Outages have a mean of 10 seconds with a minimum and maximum of 0.09 and 21 seconds. The MonthlyCharge is the average monthly charge for the customer. The monthly charge ranges from $79.97 to $290.16, with an average of $172.62. Bandwidth_GB_Year is the total amount of data, in gigabytes, used by the customer in a year. The amount of bandwidth used per year has a minimum of 155 gigs and a maximum of 7158 gigs. The average usage is 3392 gigs per year.

The discrete numeric variables are Population, Children, Age, Email, Contacts, Yearly_equip_failure. Population is the number of people within a mile radius of a customer. The average is 9756 with a minimum of 0 and a maximum of 111850. Children are the number of children in the household at the time of signup. The maximum is 10 with a minimum of 0 and a mean of 2. Age is at the time of enrollment. The average age of customers is 53 years, with a minimum of 18 and a maximum of 89. Email counts the number of emails sent to the company in the last year. The count ranges from 1 to 23, with a mean of 12. Contacts are the number of times the customer contacted technical support. Contacts range from 0 to 7, with an average of less than 1. Yearly_equip_failure counts the number of times customer's equipment has failed and replaced in the past year. The maximum is 6 with both mean and minimum of 0.

The categorical independent variables are: Area(rural, urban, suburban), Marital, Gender, Churn(ended service in the last month), Techie, Contract(month-to-month, one year, two year), Port_modem(portable modem), Tablet, InternetService(DSL, fiber optic, none), Phone, Multiple(multiple lines), OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod.

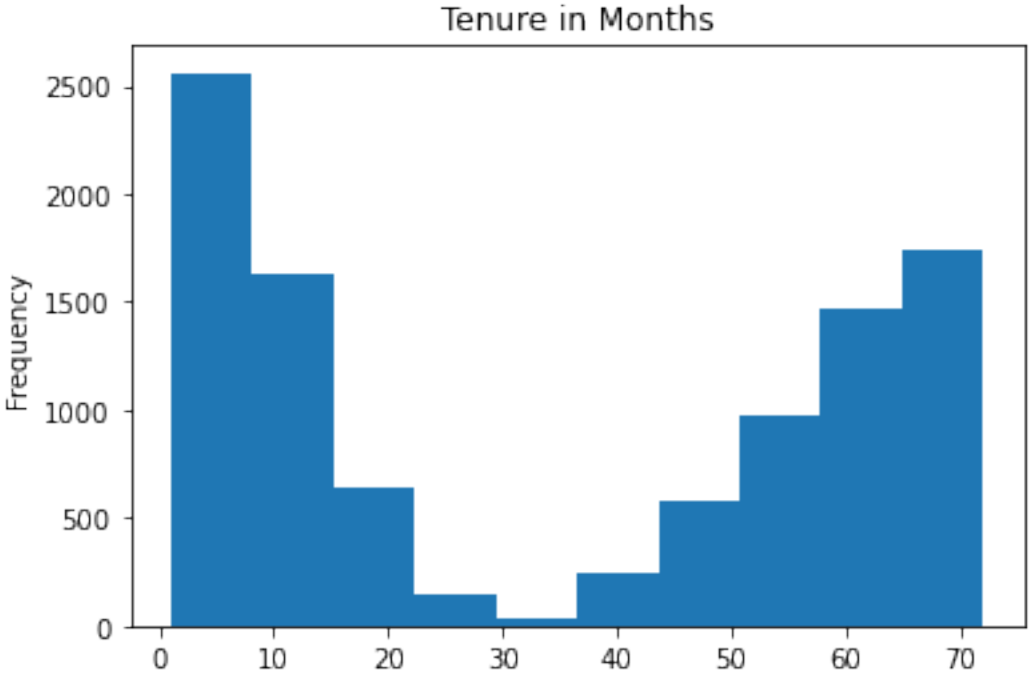
Items 1 through 8 are ordinal values. Customers' responses on a survey of which factors are most important to them, with one being most important and eight being least important. The features included timely response, timely fixes, timely replacements, reliability, options, respectful response, courteous exchange, and evidence of active listening.

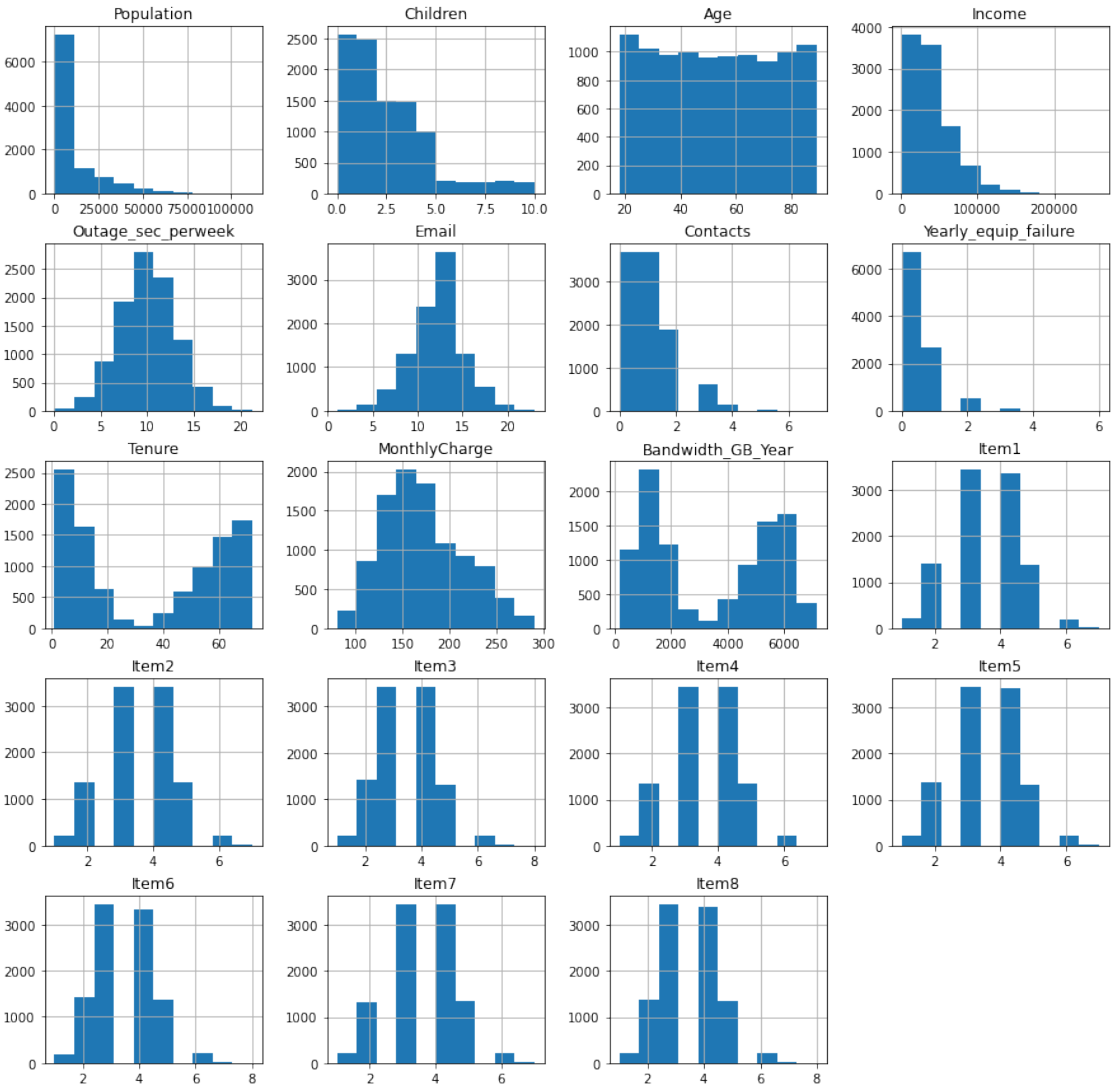
3. After loading the file, these steps used to prepare for analysis were:
   - Use .head() to view the first five rows to ensure the data set loaded and get a glance at the information it holds.
   - Use .shape() to see the number of columns and rows.
   - Use .info() to see all the column names and the data types.
   - Use .drop to remove CaseOrder, Customer_id, Interaction, UID, Lat, Lng, TimeZone, Job, City, County
   - Look at zip code and state to decide if they should be included or dropped. Use .nunique() to see the total number of zip codes and states included in the data set. There are 8583 unique zip codes and 52 states. Including these variables will make the regression equation too unwieldy to glean any information. State and Zip were dropped.
   - Check for missing values using .isnull().count().
   - Check for duplications using .duplicated()

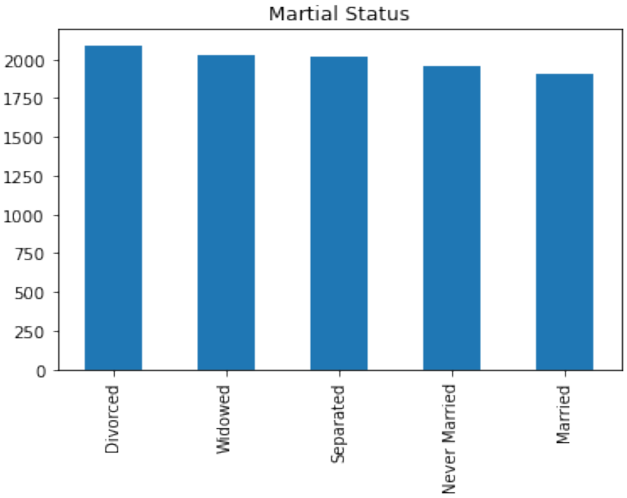
- Use .describe() to look for outliers in the numeric variables. Compare the number of unique entries in each categorical variable with the provided Data Dictionary to check for the correct number of distinct responses.
- Create dummy variable using .get_dummies on Area, Marital, Gender, Churn, Techie, Contract, Port_modem, Tablet, InternetService, Phone, Multiple, OnlineSecurity, OnlineBackup, DeviceProtection, TechSupport, StreamingTV, StreamingMovies, PaperlessBilling, PaymentMethod. Use .drop and .merge to remove original columns and add dummy to the data set. (Langford, 2017)
- Change all data types to float using .astype()

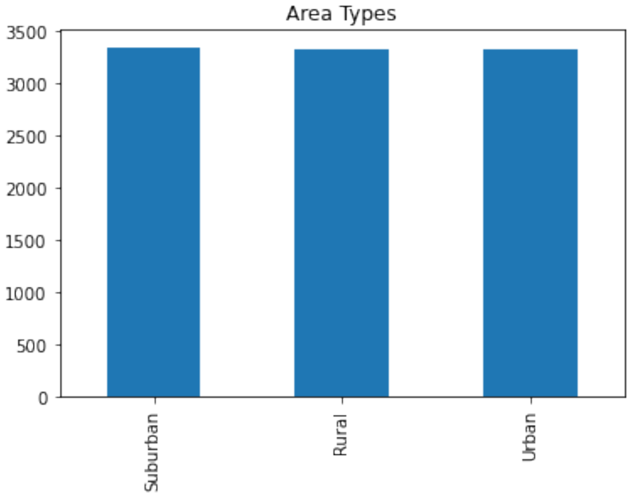
4. Univariate Visualizations

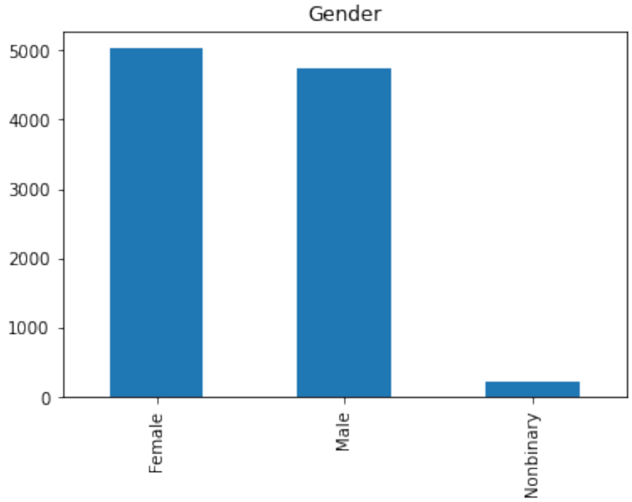
Target Variable:

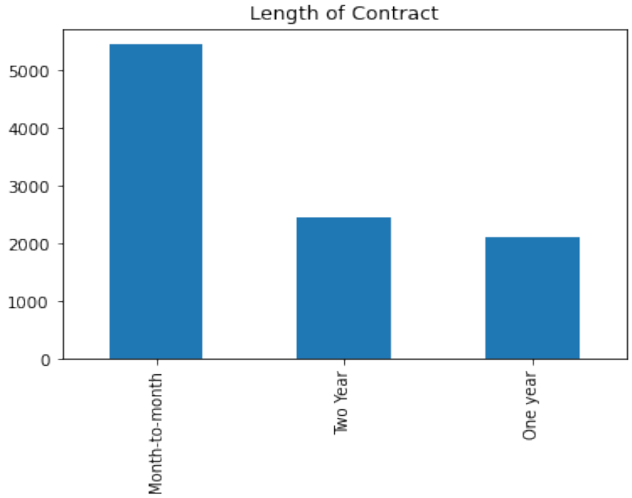


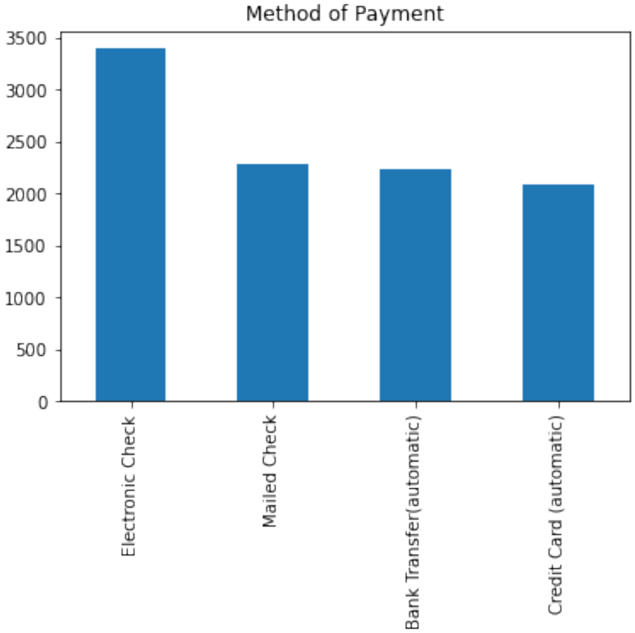

Population
0
2000
4000
6000
0 25000 50000 75000 100000
Children
0
500
1000
1500
2000
2500
0.0 2.5 5.0 7.5 10.0
Age
0
200
400
600
800
1000
20 40 60 80
Income
0
1000
2000
3000
4000
0 100000 200000
Outage_sec_perweek
0
500
1000
1500
2000
2500
0 5 10 15 20
Email
0
1000
2000
3000
0 5 10 15 20
Contacts
0
1000
2000
3000
0 2 4 6
Yearly_equip_failure
0
2000
4000
6000
0 2 4 6
Tenure
0
500
1000
1500
2000
2500
0 20 40 60
MonthlyCharge
0
500
1000
1500
2000
100 150 200 250 300
Bandwidth_GB_Year
0
500
1000
1500
2000
0 2000 4000 6000
Item1
0
1000
2000
3000
2 4 6
Item2
0
1000
2000
3000
2 4 6
Item3
0
1000
2000
3000
2 4 6 8
Item4
0
1000
2000
3000
2 4 6
Item5
0
1000
2000
3000
2 4 6
Item6
0
1000
2000
3000
2 4 6 8
Item7
0
1000
2000
3000
2 4 6
Item8
0
1000
2000
3000
2 4 6 8

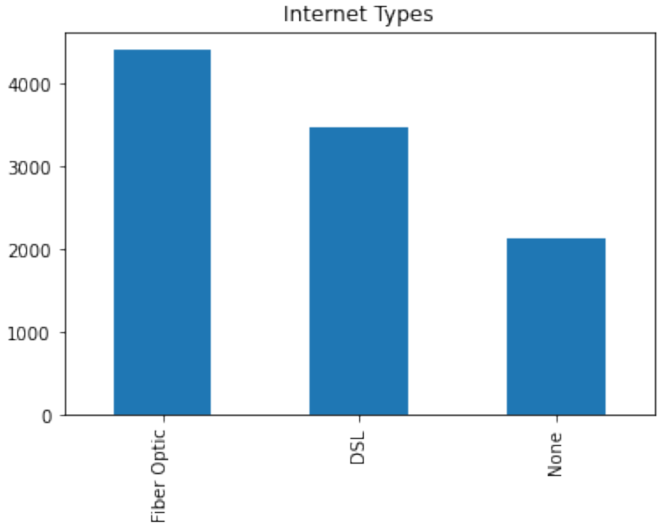

Martial Status
2000
1750
1500
1250
1000
750
500
250
0
Divorced
Widowed
Separated
Never Married
Married


Area Types
3500
3000
2500
2000
1500
1000
500
0
Suburban
Rural
Urban

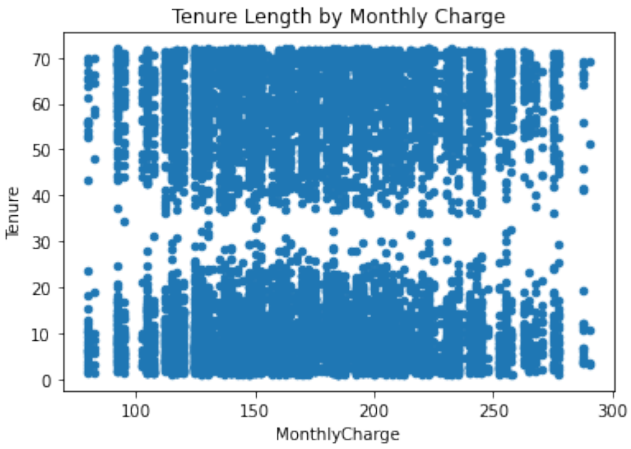

Gender
5000
4000
3000
2000
1000
0
Female
Male
Nonbinary

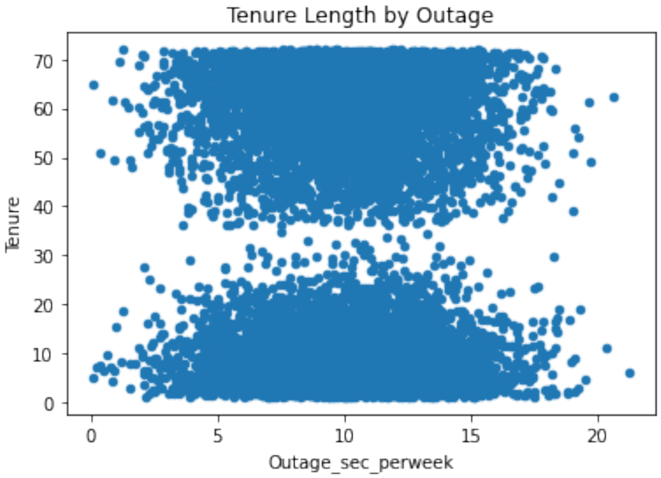

Length of Contract
5000
4000
3000
2000
1000
0
Month-to-month
Two Year
One year

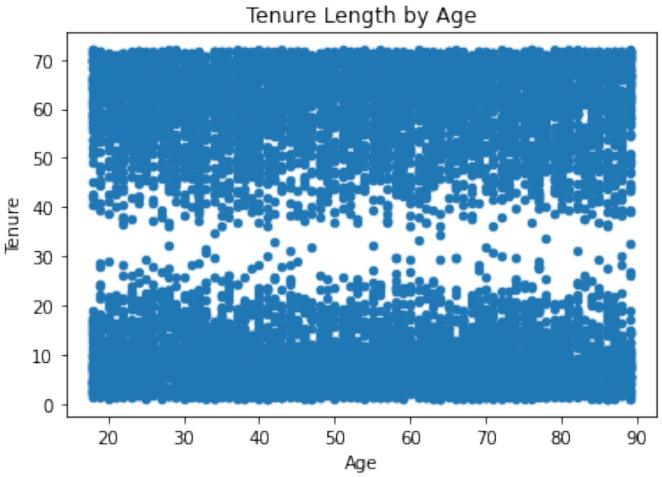

Method of Payment
3500
3000
2500
2000
1500
1000
500
0
Electronic Check
Mailed Check
Bank Transfer(automatic)
Credit Card (automatic)

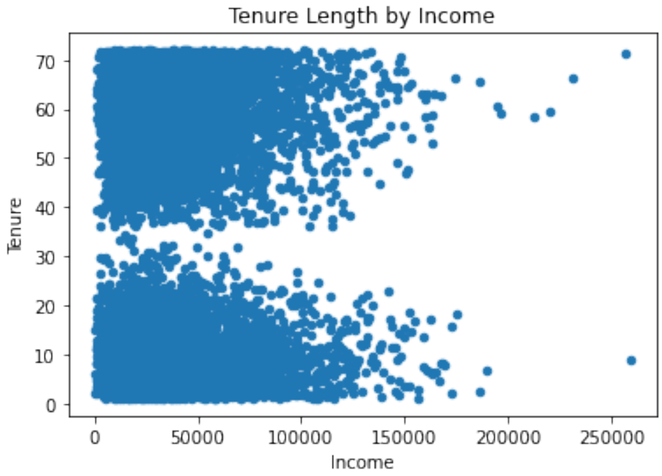

Internet Types
4000
3000
2000
1000
0
Fiber Optic
DSL
None

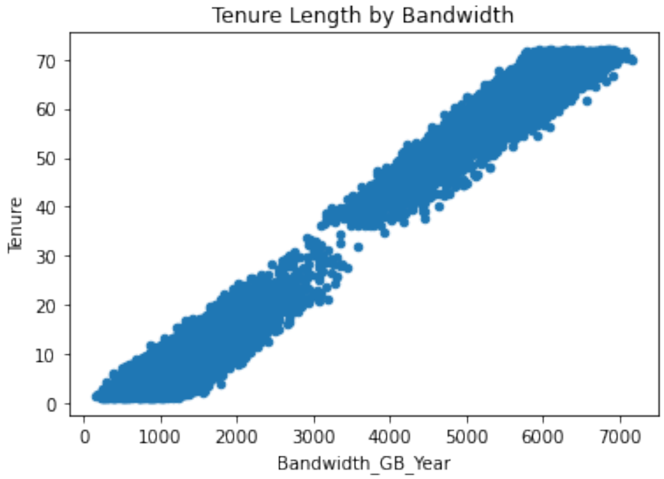
Bivariate Visualizations

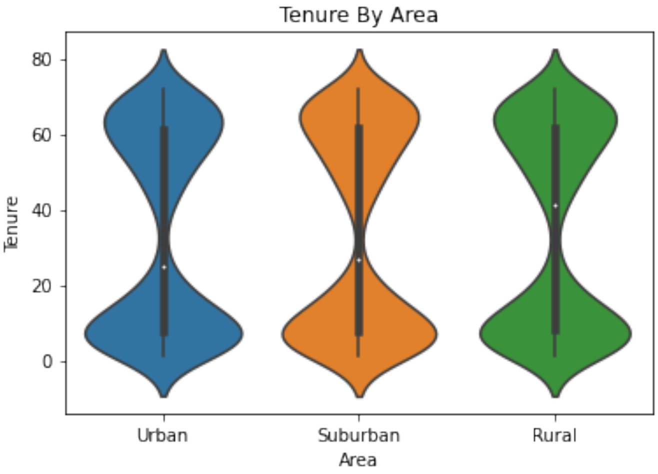

Tenure Length by Monthly Charge

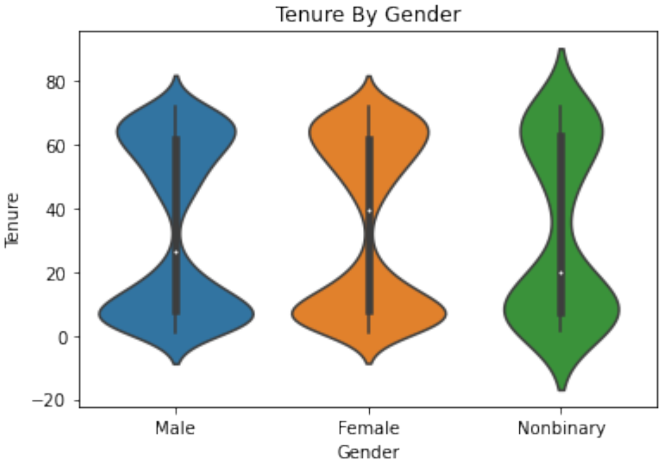

Tenure Length by Outage

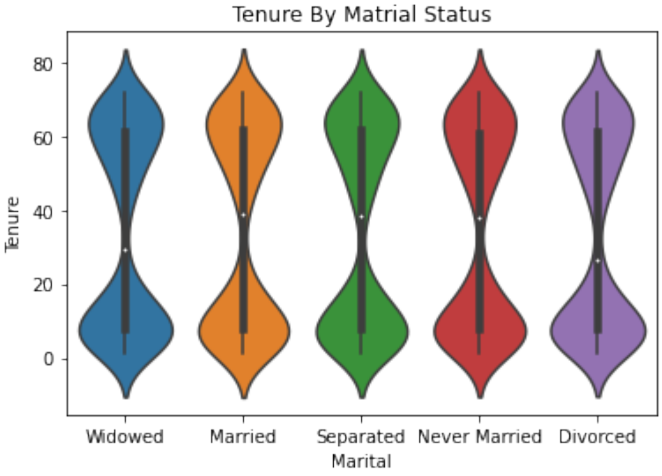

Tenure Length by Age

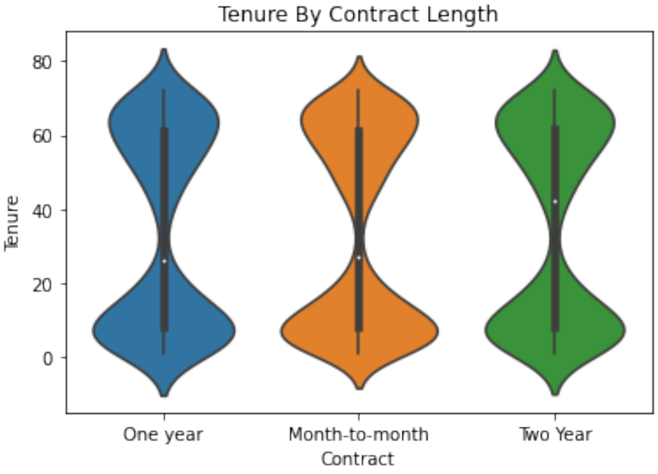

Tenure Length by Income

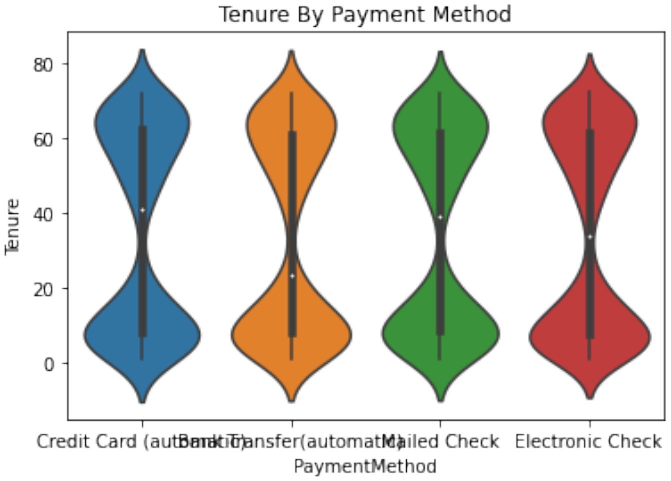

Tenure Length by Bandwidth

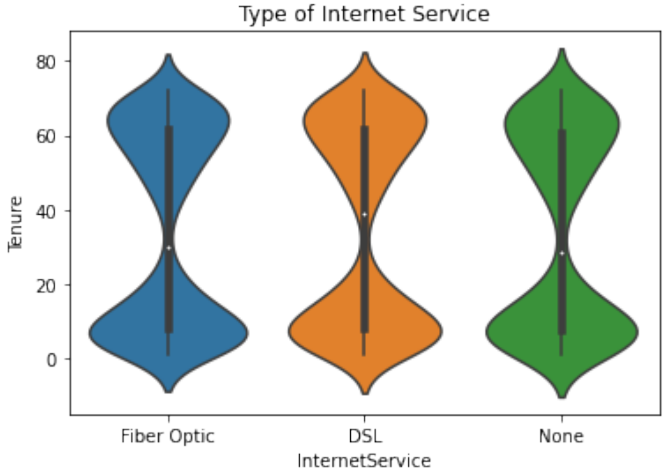

Tenure By Area


Tenure By Gender


Tenure By Matrial Status


Tenure By Contract Length


Tenure By Payment Method

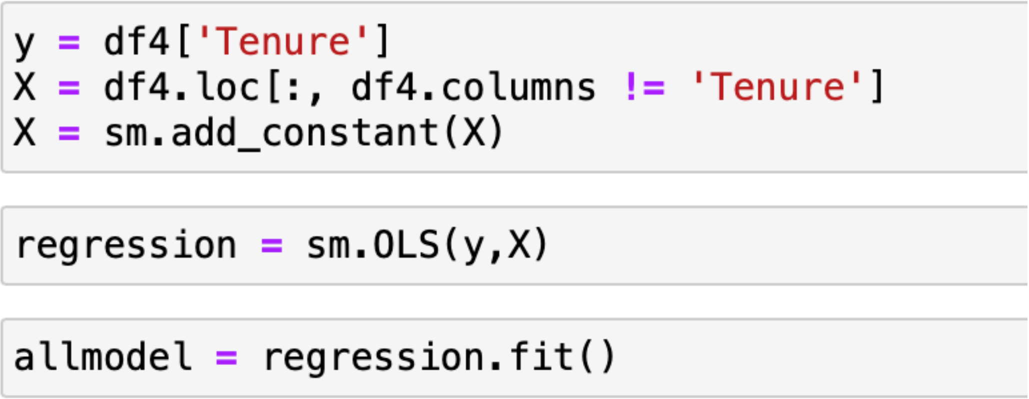

Type of Internet Service

5. Copy of prepared data attached as prepared_churn.csv

Part IV: Model Comparison and Analysis

Model Comparison

1. The model was constructed using Statsmodels and least squares regression. The code is pictured here:

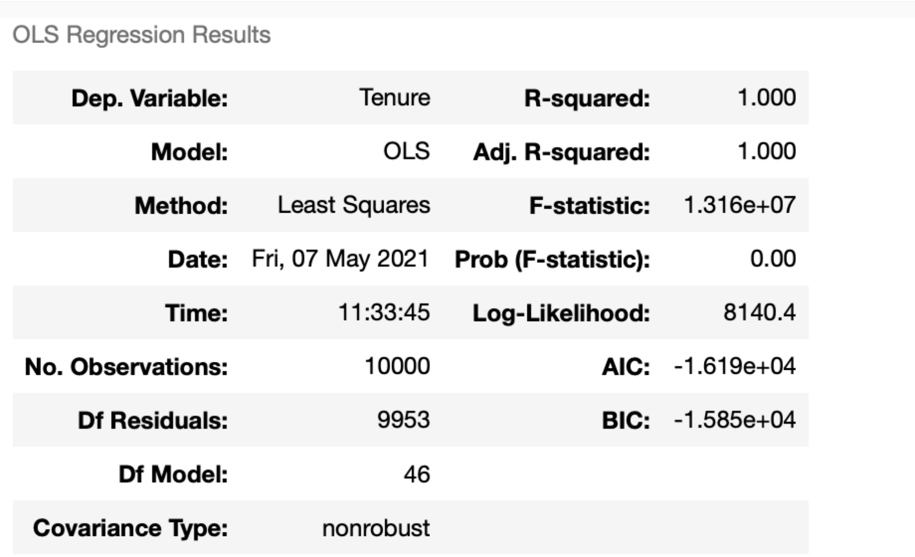


```
y = df4['Tenure']
X = df4.loc[:, df4.columns != 'Tenure']
X = sm.add_constant(X)
```

```
regression = sm.OLS(y,X)
```

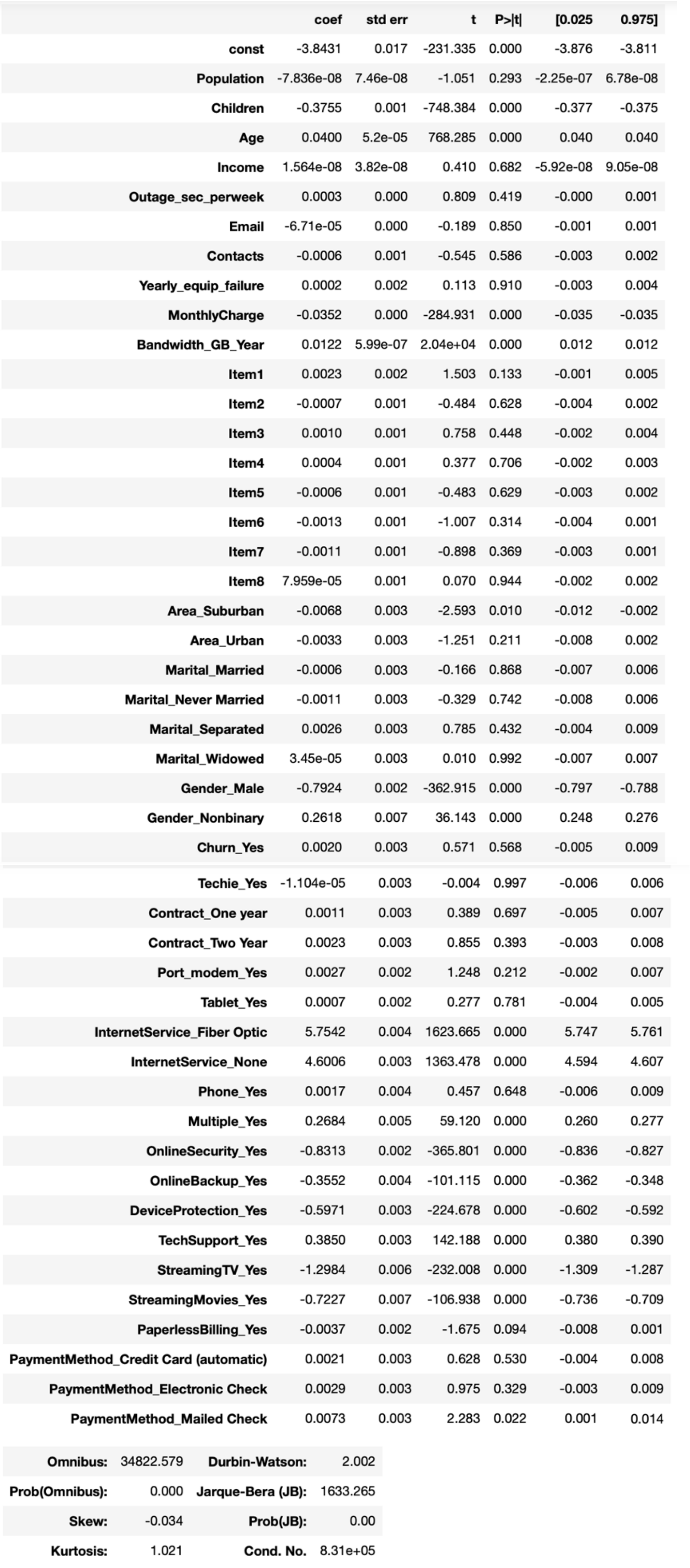
```
allmodel = regression.fit()
```

The following is the summary from the model:



OLS Regression Results

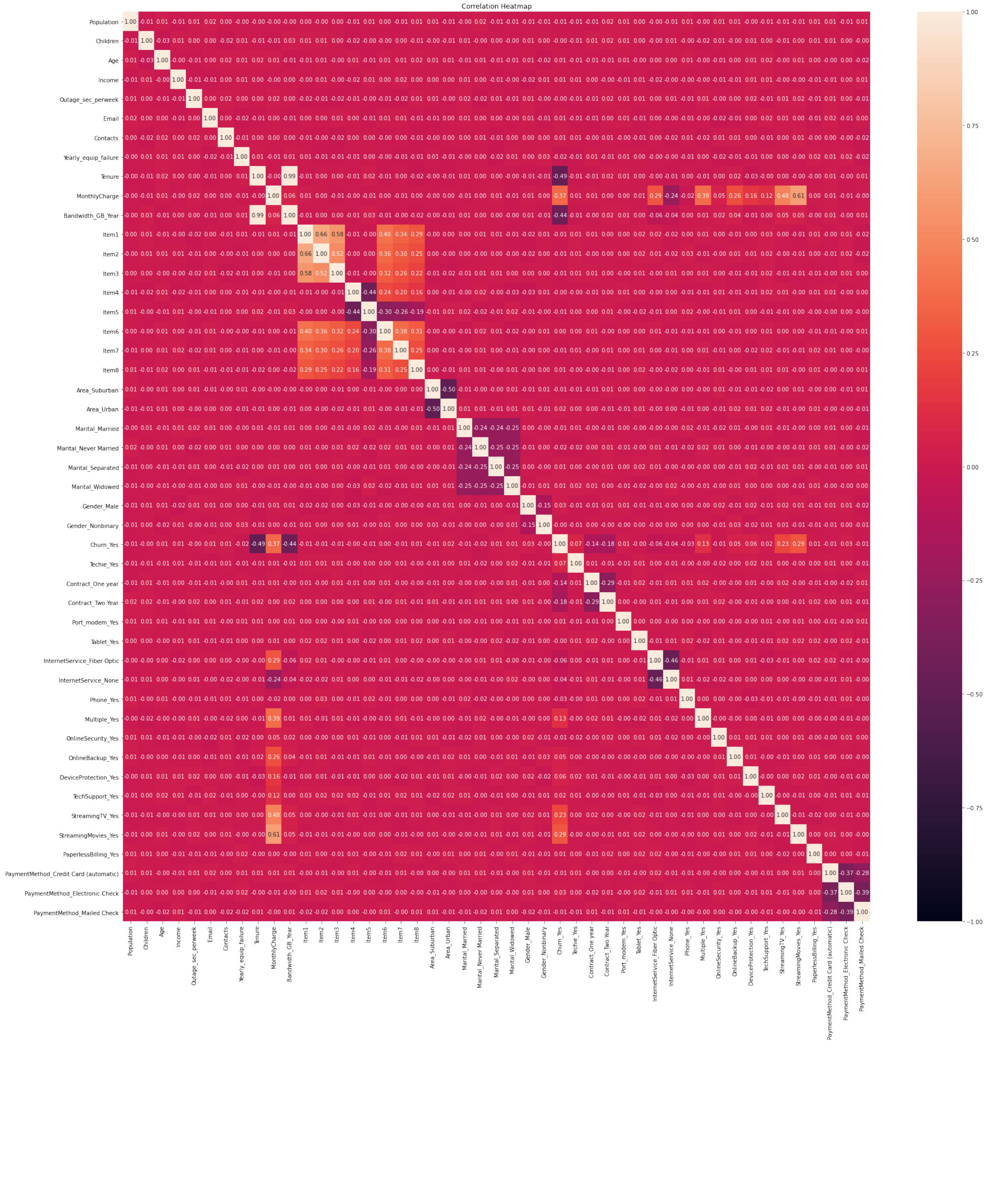
Dep. Variable:	Tenure	**R-squared:**	1.000
Model:	OLS	**Adj. R-squared:**	1.000
Method:	Least Squares	**F-statistic:**	1.316e+07
Date:	Fri, 07 May 2021	**Prob (F-statistic):**	0.00
Time:	11:33:45	**Log-Likelihood:**	8140.4
No. Observations:	10000	**AIC:**	-1.619e+04
Df Residuals:	9953	**BIC:**	-1.585e+04
Df Model:	46		
Covariance Type:	nonrobust		



	coef	std err	t	P>|t|	[0.025	0.975]
const	-3.8431	0.017	-231.335	0.000	-3.876	-3.811
Population	-7.836e-08	7.46e-08	-1.051	0.293	-2.25e-07	6.78e-08
Children	-0.3755	0.001	-748.384	0.000	-0.377	-0.375
Age	0.0400	5.2e-05	768.285	0.000	0.040	0.040
Income	1.564e-08	3.82e-08	0.410	0.682	-5.92e-08	9.05e-08
Outage_sec_perweek	0.0003	0.000	0.809	0.419	-0.000	0.001
Email	-6.71e-05	0.000	-0.189	0.850	-0.001	0.001
Contacts	-0.0006	0.001	-0.545	0.586	-0.003	0.002
Yearly_equip_failure	0.0002	0.002	0.113	0.910	-0.003	0.004
MonthlyCharge	-0.0352	0.000	-284.931	0.000	-0.035	-0.035
Bandwidth_GB_Year	0.0122	5.99e-07	2.04e+04	0.000	0.012	0.012
Item1	0.0023	0.002	1.503	0.133	-0.001	0.005
Item2	-0.0007	0.001	-0.484	0.628	-0.004	0.002
Item3	0.0010	0.001	0.758	0.448	-0.002	0.004
Item4	0.0004	0.001	0.377	0.706	-0.002	0.003
Item5	-0.0006	0.001	-0.483	0.629	-0.003	0.002
Item6	-0.0013	0.001	-1.007	0.314	-0.004	0.001
Item7	-0.0011	0.001	-0.898	0.369	-0.003	0.001
Item8	7.959e-05	0.001	0.070	0.944	-0.002	0.002
Area_Suburban	-0.0068	0.003	-2.593	0.010	-0.012	-0.002
Area_Urban	-0.0033	0.003	-1.251	0.211	-0.008	0.002
Marital_Married	-0.0006	0.003	-0.166	0.868	-0.007	0.006
Marital_Never Married	-0.0011	0.003	-0.329	0.742	-0.008	0.006
Marital_Separated	0.0026	0.003	0.785	0.432	-0.004	0.009
Marital_Widowed	3.45e-05	0.003	0.010	0.992	-0.007	0.007
Gender_Male	-0.7924	0.002	-362.915	0.000	-0.797	-0.788
Gender_Nonbinary	0.2618	0.007	36.143	0.000	0.248	0.276
Churn_Yes	0.0020	0.003	0.571	0.568	-0.005	0.009
Techie_Yes	-1.104e-05	0.003	-0.004	0.997	-0.006	0.006
Contract_One year	0.0011	0.003	0.389	0.697	-0.005	0.007
Contract_Two Year	0.0023	0.003	0.855	0.393	-0.003	0.008
Port_modem_Yes	0.0027	0.002	1.248	0.212	-0.002	0.007
Tablet_Yes	0.0007	0.002	0.277	0.781	-0.004	0.005
InternetService_Fiber Optic	5.7542	0.004	1623.665	0.000	5.747	5.761
InternetService_None	4.6006	0.003	1363.478	0.000	4.594	4.607
Phone_Yes	0.0017	0.004	0.457	0.648	-0.006	0.009
Multiple_Yes	0.2684	0.005	59.120	0.000	0.260	0.277
OnlineSecurity_Yes	-0.8313	0.002	-365.801	0.000	-0.836	-0.827
OnlineBackup_Yes	-0.3552	0.004	-101.115	0.000	-0.362	-0.348
DeviceProtection_Yes	-0.5971	0.003	-224.678	0.000	-0.602	-0.592
TechSupport_Yes	0.3850	0.003	142.188	0.000	0.380	0.390
StreamingTV_Yes	-1.2984	0.006	-232.008	0.000	-1.309	-1.287
StreamingMovies_Yes	-0.7227	0.007	-106.938	0.000	-0.736	-0.709
PaperlessBilling_Yes	-0.0037	0.002	-1.675	0.094	-0.008	0.001
PaymentMethod_Credit Card (automatic)	0.0021	0.003	0.628	0.530	-0.004	0.008
PaymentMethod_Electronic Check	0.0029	0.003	0.975	0.329	-0.003	0.009
PaymentMethod_Mailed Check	0.0073	0.003	2.283	0.022	0.001	0.014

Omnibus:	34822.579	**Durbin-Watson:**	2.002
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	1633.265
Skew:	-0.034	**Prob(JB):**	0.00
Kurtosis:	1.021	**Cond. No.**	8.31e+05

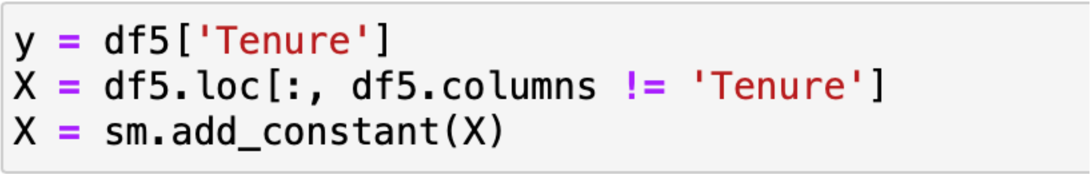
2.	In this model, the adjusted R-squared value was one, and that the model explains all the variance. The model is a perfect fit. All the variance is explained by the model. Given that there are so many variables, the model is probability overfitted. This reduces the model's usefulness for generalization. It will not be able to make predictions based on new data. Another point to look at is the p-values. If they are larger than 0.05, they are not significant and can be removed. The third warning sign is the condition number test, Cond. No. in the summary. If it is over 30, there is instability in the model due to multicollinearity. The condition number test for this model is 831000, well over the threshold. We can use a heatmap to look for instances of high correlation between variables. (Massaron & Boschetti, 2016)



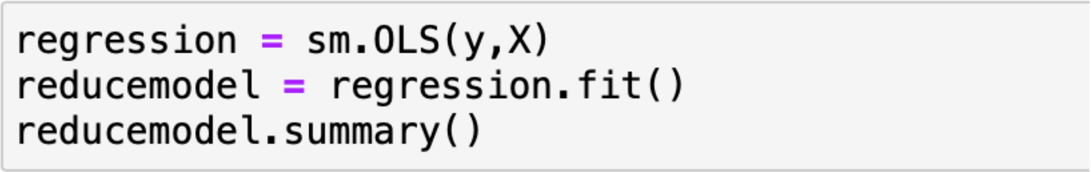
There are a few variables that are correlated to each other at a higher degree than others. With so many variables, it would not be efficient to remove each one at a time to test the effect. The Lasso method was used to reduce the number of variables in the model. This method was chosen because it is faster than a wrapper method like RFE. Lasso is more accurate than a filter method, like using just correlation values. It looks at all the features at once and is not prone to overfitting. (Pykes, 2020)

After using the Lasso method, these are the remaining variables: MonthlyCharge, Bandwidth_GB_year, Churn_Yes, and InternetService_Fiber Optic.

3. A new model was built with the above-listed variables.

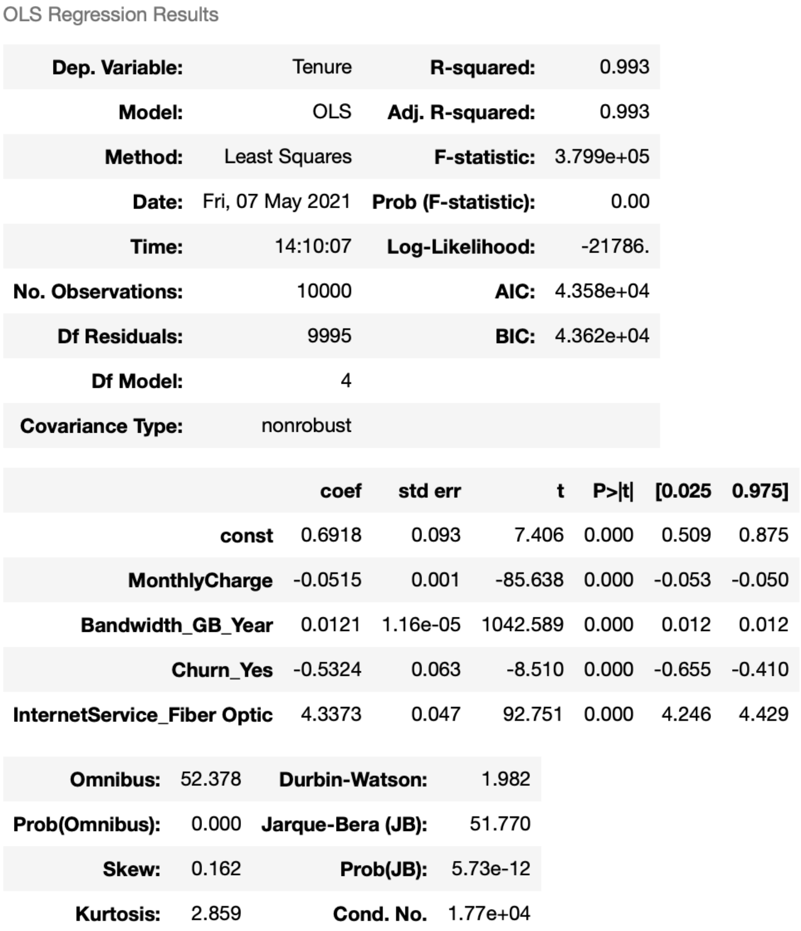


```
y = df5['Tenure']
X = df5.loc[:, df5.columns != 'Tenure']
X = sm.add_constant(X)
```



```
regression = sm.OLS(y,X)
reducemodel = regression.fit()
reducemodel.summary()
```

The r



OLS Regression Results

Dep. Variable:	Tenure	**R-squared:**	0.993
Model:	OLS	**Adj. R-squared:**	0.993
Method:	Least Squares	**F-statistic:**	3.799e+05
Date:	Fri, 07 May 2021	**Prob (F-statistic):**	0.00
Time:	14:10:07	**Log-Likelihood:**	-21786.
No. Observations:	10000	**AIC:**	4.358e+04
Df Residuals:	9995	**BIC:**	4.362e+04
Df Model:	4		
Covariance Type:	nonrobust		

	coef	**std err**	**t**	**P>|t|**	**[0.025**	**0.975]**
const	0.6918	0.093	7.406	0.000	0.509	0.875
MonthlyCharge	-0.0515	0.001	-85.638	0.000	-0.053	-0.050
Bandwidth_GB_Year	0.0121	1.16e-05	1042.589	0.000	0.012	0.012
Churn_Yes	-0.5324	0.063	-8.510	0.000	-0.655	-0.410
InternetService_Fiber Optic	4.3373	0.047	92.751	0.000	4.246	4.429

Omnibus:	52.378	**Durbin-Watson:**	1.982
Prob(Omnibus):	0.000	**Jarque-Bera (JB):**	51.770
Skew:	0.162	**Prob(JB):**	5.73e-12
Kurtosis:	2.859	**Cond. No.**	1.77e+04

The reduced model now has an adjusted R-squared value of .993, a reduction of .007. All the p values are 0, below the limit of 0.05. The condition number is still over 30 but was reduced to 17700, a drop of 813300. This new model is slightly better than the original.

Model Analysis

1. The focus of this multiple linear regression model was Tenure. The goal was to find, given the current dataset, a model to predict a customer's Tenure. A model was built using all the predictors and the statsmodel's least squares regression function. The resulting model had an R-squared of 1.00. The model is a perfect fit and most likely not possible in a real-world situation. A closer look at the p-values for the predictive variables showed that many were greater than 0.05 and are not statistically significant. Also, the condition number was 831000 with is well over the threshold of 30. A large condition number indicates that there is collinearity between the predictive variables.

An evaluation of a correlation heatmap showed that some variables moderately correlated but not so large that one variable could be dropped. The lasso method was then used to reduce the number of variables in the model. This method was used because it is fast, accurate, and is not prone to overfitting. The lasso method returned four variables: MonthlyCharge, Bandwidth_GB_Year, Chrun_Yes, InternetService_Fiber Optic.

A second model was built using the above variables. The reduced model returned an R-squared of 0.993, which is slightly more realistic. This model explains 99.3% of the variability in the variables. The p-values for the predictor variables are all 0.00. The condition number was reduced to 17700.

2. The output from the initial model is shown in Figure 1. The output from the reduced model is shown in Figure 2. The residual error calculations are in the tables below.

Initial Model	
Mean Absolute Error	0.107
Mean Squared Error	0.011
Root Mean Squared Error	0.107

Reduced Model	
Mean Absolute Error	1.706
Mean Squared Error	4.569
Root Mean Squared Error	2.138

The errors in the initial model are very small and not very realistic. The errors for the reduced model are better.

The residual plots showed the reduced model (Figure 4) improves the initial model (Figure 3).

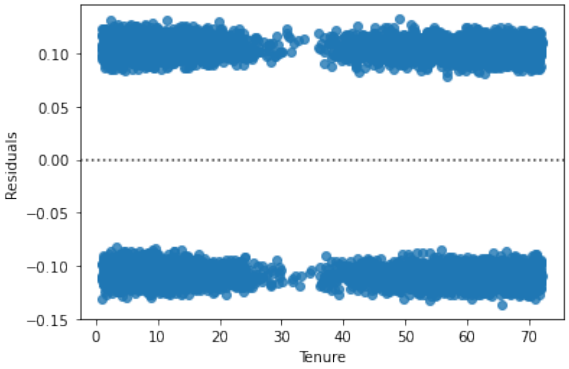


Figure 3

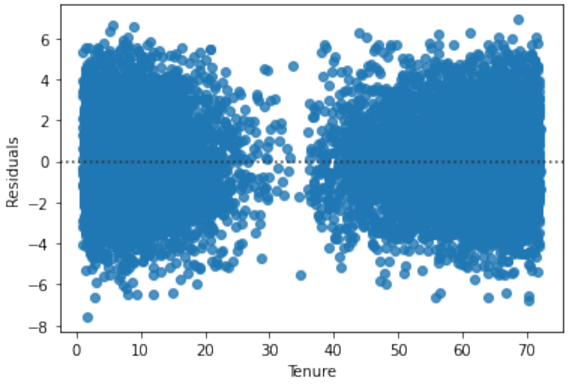


Figure 4

Part V: Data Summary and Implications

1. The equation for the reduced model is:

$$y = 0.6918 - 0.0515(\text{MonthlyCharge}) + 0.0121(\text{Bandwidth\_GB\_Year}) - 0.5324(\text{Chrun\_Yes}) + 4.3373(\text{InternetService\_Fiber Optic})$$

MonthlyCharge and Churn_Yes have a negative effect on the Tenure of customers. InternetService_Fiber Optic and Bandwidth_GB_Year has a positive impact on the Tenure of customers. InternetService_Fiber Optic has the most significant influence, with a coefficient of 4.3373.

There are some limitations to this analysis. The data set represents a tiny portion of the total population. A quick Google search showed that three of the largest telecommunication companies have 20 million customers each. The data set with 10,000 records is not very large in comparison.

The time frame is limited. The needs of customers quickly change as technology grows. Looking at this narrow timeframe may not be enough to show the growing demands of customers.

There are some assumptions needed for multiple linear regression that does not hold. First, the data needs to be normally distributed. Looking at the univariate graph, it is easy to see that most of the predictive variables do not meet this assumption. Next, the relationship needs to be linear. This does not hold for most of the predictor variables. Some of the variables are binary and would function better in logistic regression. Finally, the residual errors after fitting the model needs to be evenly distributed across the graph. Figure 3 and Figure 4 shows the distribution of errors are not evenly distributed for either model. The reduced model is slightly better than the initial model, but there is a hole in the middle for the data.

2. Based on the equation from the reduced model, it is recommended that the company expand fiber optic internet service to current and future customers. The increase in the monthly cost will only slightly decrease the positive effect of fiber optic internet service.

Works Cited

Langford, R. (2017, March 24). *The Dummy's Guide to Creating Dummy Variables*. Retrieved April 2021, from towards data science: https://towardsdatascience.com/the-dummys-guide-to-creating-dummy-variables-f21faddb1d40

Massaron, L., & Boschetti, A. (2016). *Regression analysis with Python.* Packt Publishing.

Pykes, K. (2020). *Gettinf Started with Feature Selection*. Retrieved May 2021, from KDnuggets: https://www.kdnuggets.com/2020/08/getting-started-feature-selection.html

Using Statistical Regression Methods in Education Research. (2011, July 22). Retrieved from Restore at National Centre for Research Methods: https://www.restore.ac.uk/srme/www/fac/soc/wie/research-new/srme/modules/mod3/3/index.html