***I: Tool Selection***  
Execute data extraction from the “Customer Data” web link using data mining software (Python, R, or SAS). Provide a screen shot of the code you have written and its successful application with a copy of all the extracted data.

1. Describe the benefits of using the tool you have chosen (Python, R, or SAS) for extracting data in this scenario.

I am choosing to use R for multiple reasons: It’s open-source unlike SAS, so in a sense more important for me to familiarize myself with as I know I will always have access to it. I like its flexibility. Its package libraries and capabilities are more diverse than either those of SAS and Python, especially for graphics– and visualization is an important part of this assignment. Since I am not making my analysis results available via an app, etc. the major benefit of Python relative to R, ease of accessibility, is not a factor for me. SAS might be a more natural choice for more intensive financial or commercial analysis, but this is a small project.

1. Define the objectives or goals of the data analysis. Ensure that your objectives or goals are reasonable within the scope of the scenario and are represented in the available data.

The goals of my analysis were to isolate which demographic variables might contribute to choices to enroll in paperless billing or to discontinue service with the company, and which result in higher monthly charges. I wanted to be able to give predictions for these three factors based on the demographic variables present in the data. I was especially interested in gender as an independent variable however my analysis proved that its effects were not terribly significant so I moved on to use others over the course of the project.  
  
I was drawn to using demographic variables as predictors simply because demographic data is more an area of interest for me than is commercial data. MonthlyCharge and Churn were chosen as potential dependent variables due to their likely business value (MonthlyCharge was chosen over TotalCharge so that tenure would not need to be taken into account – to pare down what I was doing). PaperlessBilling was chosen because the relationship between environmentally conscious choices and demographics is, again, simply interesting to me.

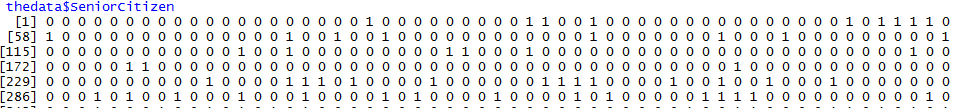
1. Select a descriptive method *and* a nondescriptive method (i.e., predictive, classification, or probabilistic techniques) you will use to analyze the data, and explain how the methods you have selected are appropriate for the objectives or goals you have defined.  
     
   I explored the data using several descriptive techniques, including basic summarization and correspondence matrix techniques, as well as Multiple Analysis of Variance, and Multiple Correspondence Analysis. I chose MANOVA and MCA because I wanted, at least initially, to look at several dependent and independent variables simultaneously, and all of the demographic variables I chose were two-level categorical (though at points I handled them as binaries).  
     
   For predictive analysis I chose to use linear and multiple multivariate regression models. I decided on them because, after looking at the data, it seemed that the relationships worth examining were clear enough in their linearity that more complex techniques like Random Forest or Decision Trees would be unnecessary if not unhelpful. I decided to use them in spite of the fact that my independent variables were all binary – more on that decision later.

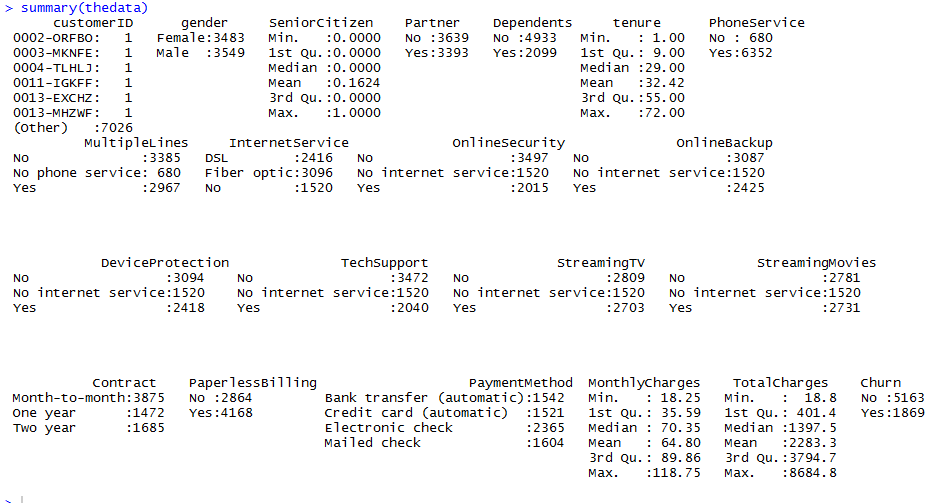
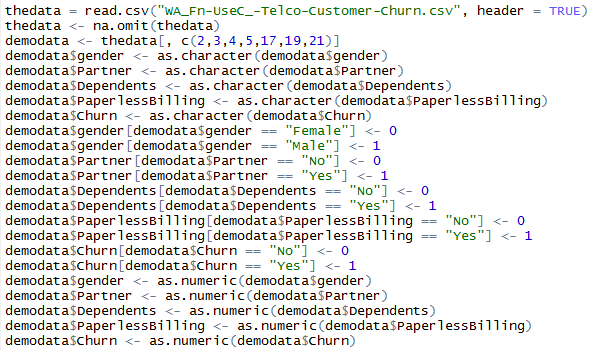
***II: Data Exploration and Preparation***  
Clean the data you have extracted and save as .xls or .xlsx format for submission. Be sure to address all necessary formatting, converting, and missing data.

1. Describe the target variable in the data and indicate the specific type of data the target variable is using, including examples that support your claims.

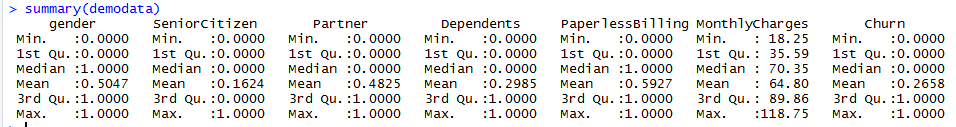
After engaging in some exploratory analysis I eventually settled on MonthlyCharges as my primary target variable. The data in this variable is continuous and numeric, representing a monetary value.  
  

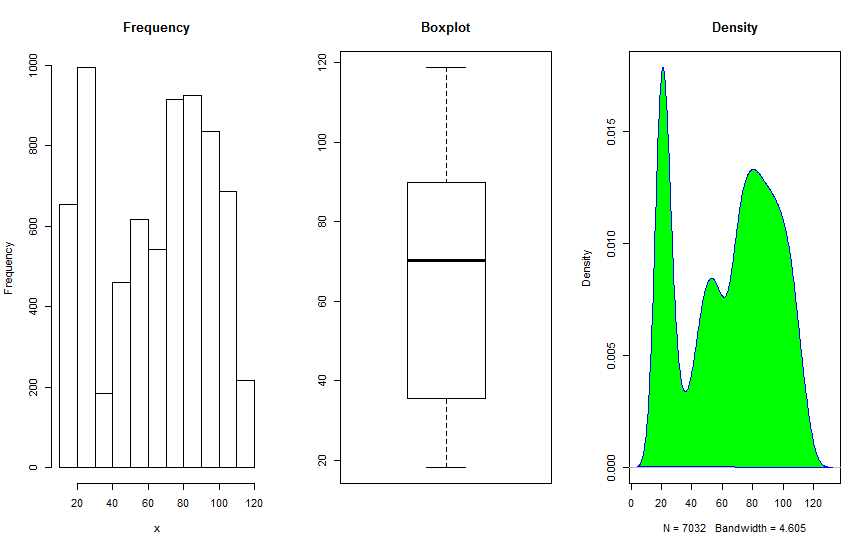
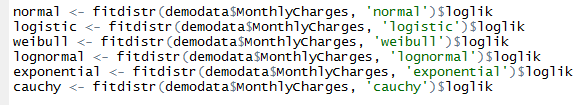
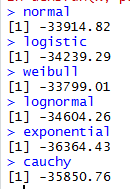
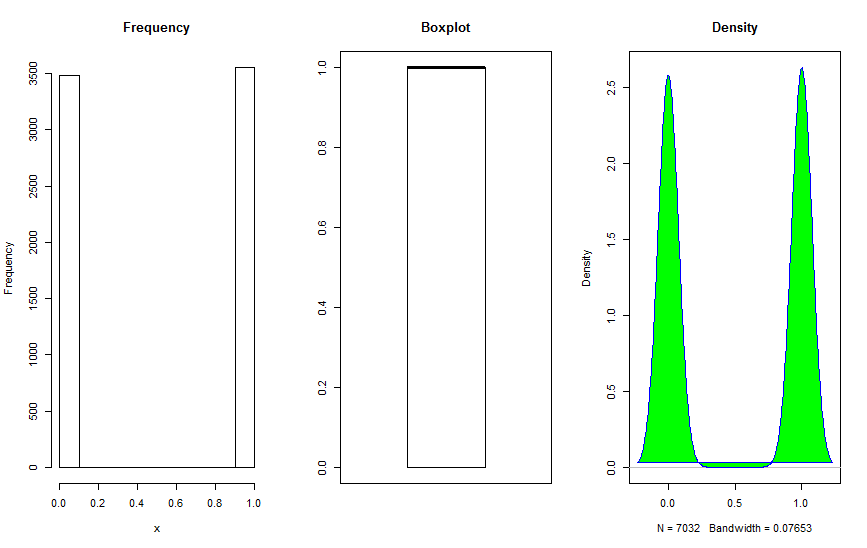

1. Describe an independent predictor variable in the data and indicate the specific type of data being described. Use examples from the data set that support your claims.

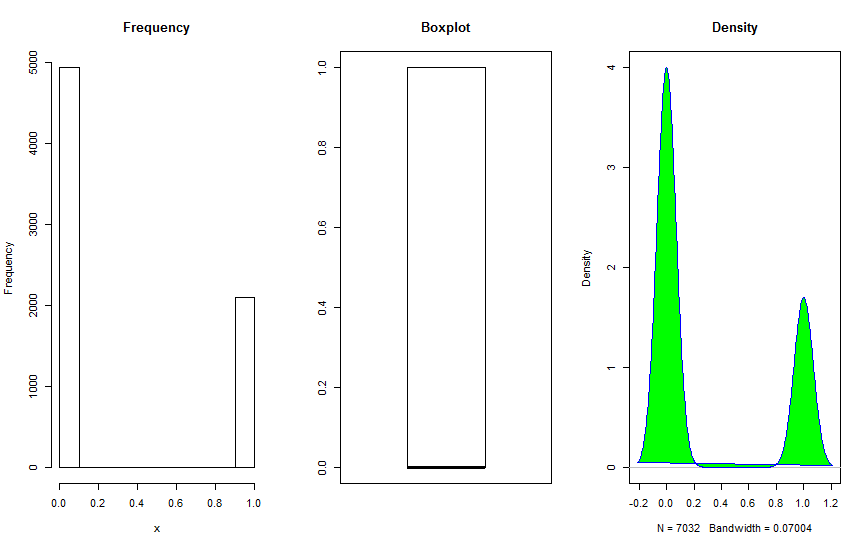
I made use of three independent variables: SeniorCitizens, Dependents and Partner. But again, after exploratory analysis, it became clear to me that SeniorCitizens would be my most significant independent variable, so here I’ll use it as an example. It reads as an integer in the initial dataset but is a Boolean with 0 representing a False value and 1 representing True.  
  


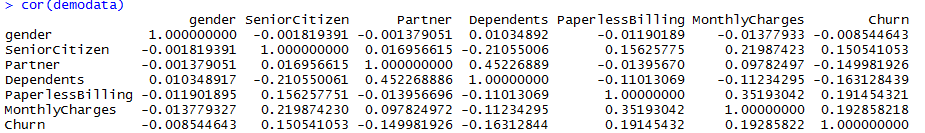
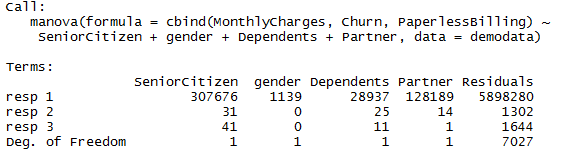
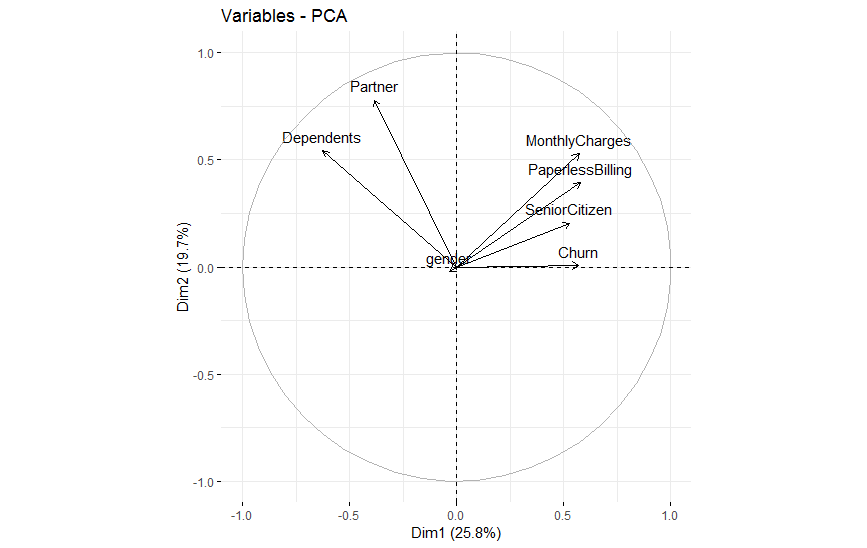
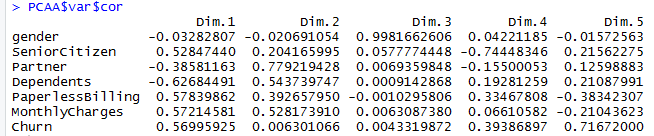
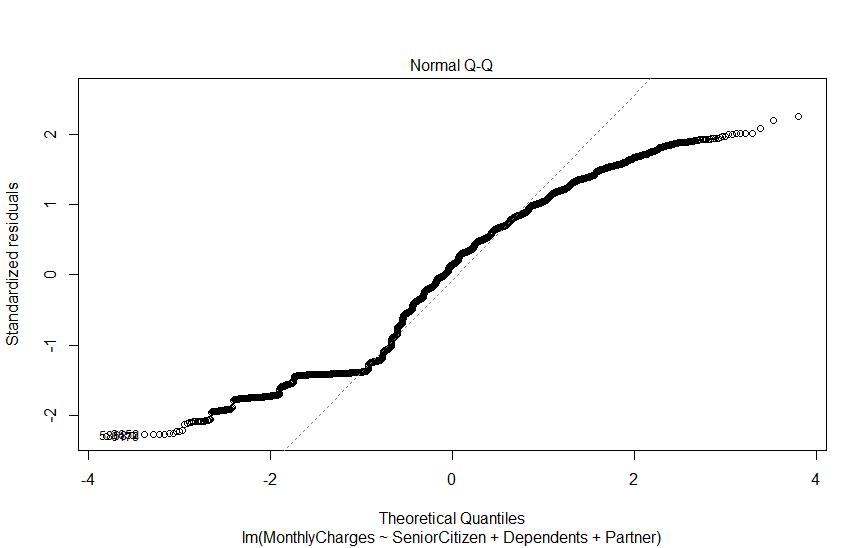
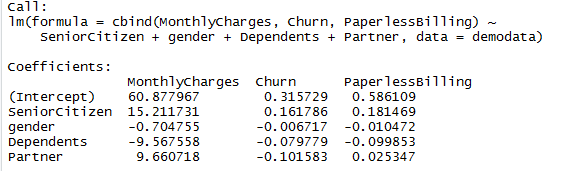
1. Propose the goal in manipulation of the data and define your data preparation aims.  
     
   My goals for manipulating and preparing the data were as follows:  
     
   Limit to three potential target variables: MonthlyCharges, Churn, and PaperlessBilling. These were chosen based on a combination of potential value of the business insights (for the first two) and my own interest (the third)  
     
   Limit potential independent variables to those dimensions of the original dataset having to do with demographics: SeniorCitizen, Partner, Dependents, and gender.   
     
   Omit rows with missing values.  
     
   Transform all categorical variables with two possible categories (so, all variables except MonthlyCharges) to binary, to match the datatype of SeniorCitizens and to make working with them easier in later stages. For gender, 0 = Female, 1 = Male. For all others, No = 0, Yes = 1.
2. Define the statistical identity of the data, including the essential criteria and phenomenon to be predicted.  
     
   The initial dataset contained 21 variables concerning a telecommunications and internet company’s customer accounts, making account holders the statistical entity being studied. It is a relatively versatile dataset and could potentially be used to do everything from assign risk ratings to customers based on their likelihood of terminating their contract to helping sales teams determine which services bring in higher monthly revenue and determine who they should target their marketing toward to determining which demographics are most likely to use customer support services or sign longer term contracts.  
     
     
     
     
   My own initial aim was to find and analyze relationships between the datasets demographic variables (SeniorCitizen, gender, Partner, Dependents) and two variables I determined to be of high value to business operations, MonthlyCharges and Churn. I also decided to look at demographic variables in relation to PaperlessBilling, which might be of interest to a company attempting to meet environmental compliance standards.   
     
   After data exploration and some preliminary modelling was completed, my aim became more specific: predict the amount of an account holder’s monthly bill based on whether they are or are not a Senior Citizen. This sort of analysis might be of interest to any business considering offering specialized contracts to Seniors, or to any nonprofit interested in ensuring that Senior Citizens have ready access to communications technologies or otherwise are getting their needs met and bills paid. If paired in analysis with other, external datasets describing things like income and additional living expenses, it might even be usable to help determine if Seniors are at risk of losing services due to financial overextension.  
     
   The most essential criteria of the undertaking would be to determine if such a relationship between Senior status and monthly charges does exist, ideally while controlling for the impacts of other demographic variables. Developing a model that successfully predicts increases or decreases in monthly charge based on an account holder’s Senior status would also qualify as an essential criteria for project success.
3. Explain the steps used to clean the data and how you addressed any anomalies or missing data.  
     
   Potential missing data was omitted using na.omit. The data was subsetted only to include demographic variables and select other variables the interactions between which I was curious about – so, gender, partner, dependents, senior citizen, paperless billing, monthly charges, and churn. Factors with only two levels – all columns but monthly charges - were converted to characters, had their levels assigned to binary values of 0 or 1, and then were converted to numeric. I did this to increase the ease with which I could experiment with various analysis methods and to increase the variety of methods at my disposal, since some methods frequently used only for continuous quantitative variables do also give valid results when applied to numerically represented true/false values. I did not find any concerning outliers in my chosen variables.  
     
   

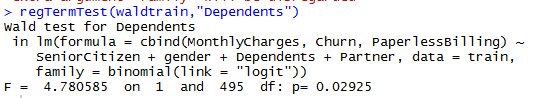
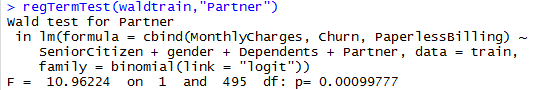
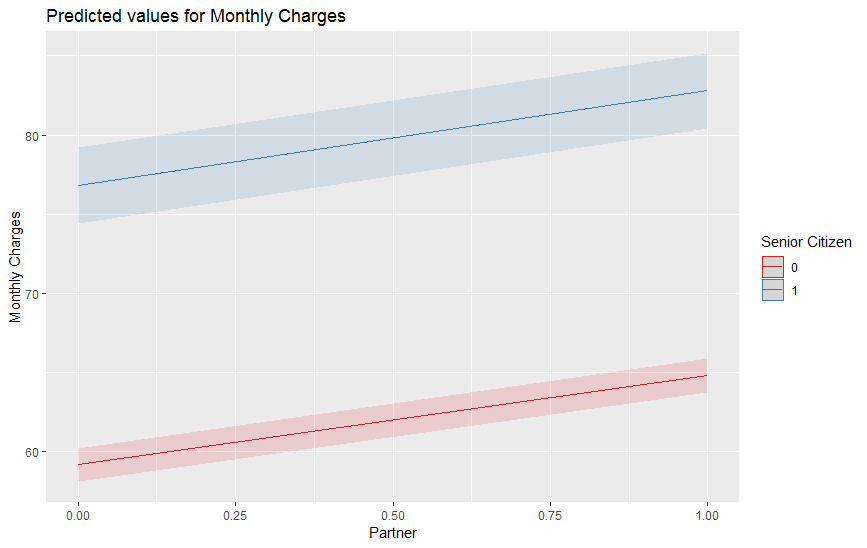
***III: Data Analysis***  
For each of the following steps, be sure to clearly indicate each step within your data sheet with a screen shot and annotations in your final submission. All algorithms used need to be clearly identified in the screen shot and submission.

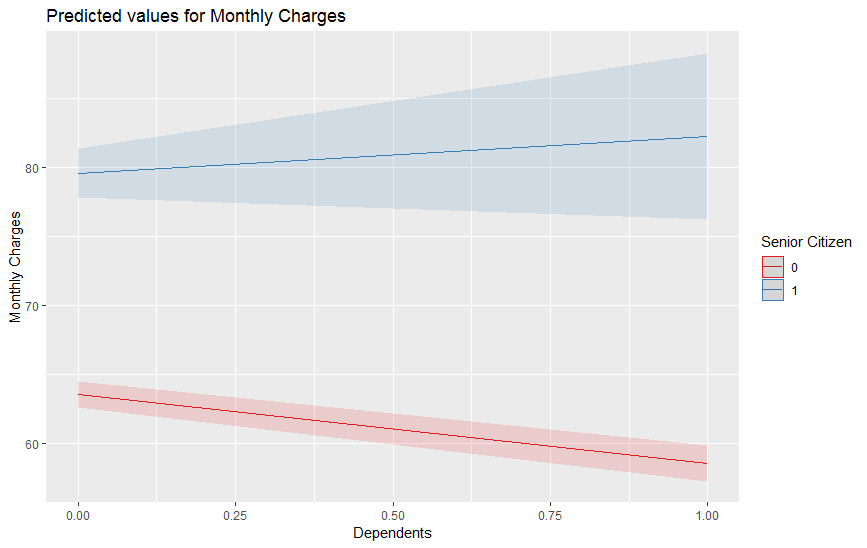
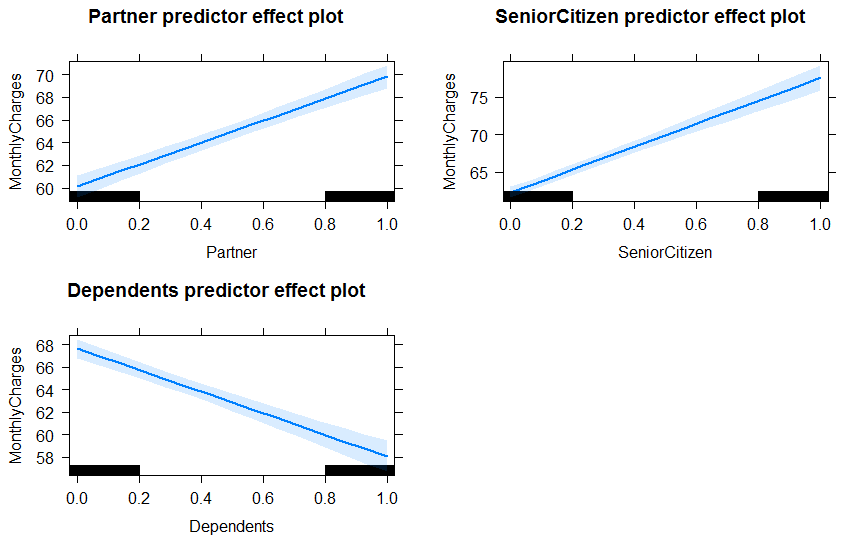
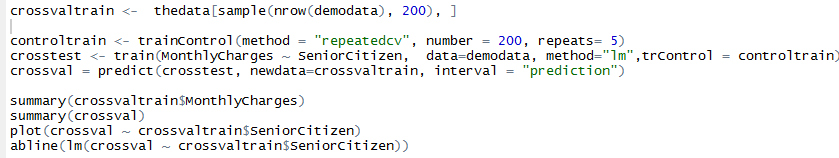
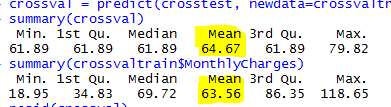
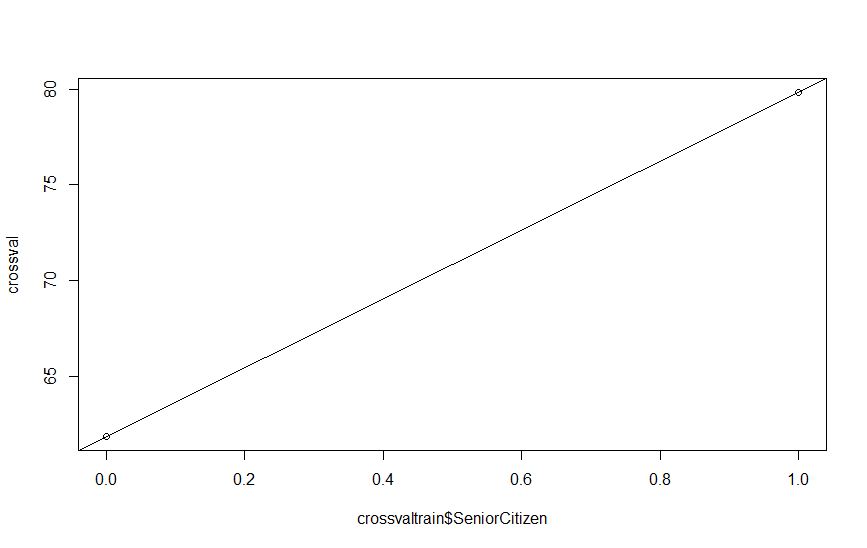
1. Identify the distribution of variables using univariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.  
     
   

I initially did a simple summary on the data to get a first impression, then plotted the only continuous variable (MonthlyCharges) using three charts that are good for visualizing distribution, side by side: a histogram, a boxplot, and a density chart.  
  
Univariate visualization for MonthlyCharges variable:  
  
After visually ruling out obviously incorrect distributions (such as Poisson) I compared the remaining possibilities using the MASS package fitDistr function’s log likelihood output.  
  
  
  
  
  
The results indicate that the Weibull distribution is the closest match, with Normal second.  
  
The remaining variables,being two-level categorical or binary, all demonstrate clear binomial or Bernoulli distributions, for example:   
Univariate visualization for Gender variable:  
  


Univariate visualization for Dependents variable:  


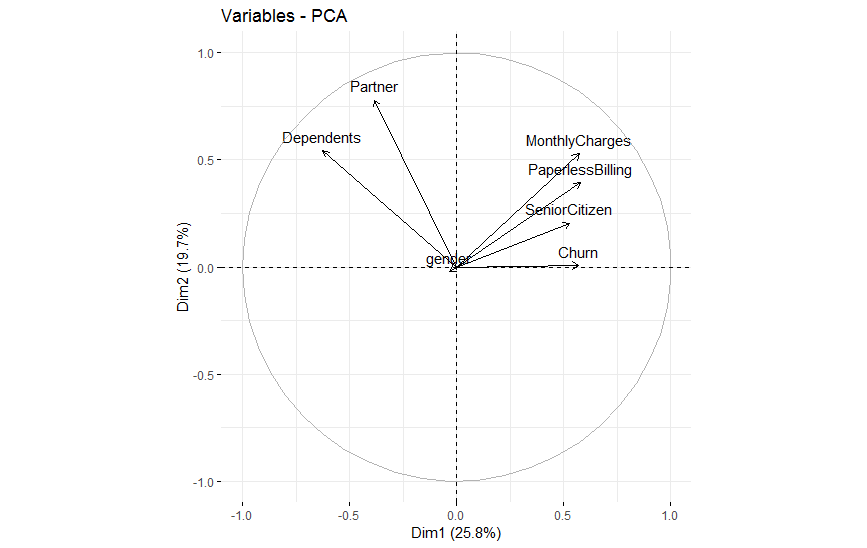
1. Identify the distribution of variables using bivariate statistics from your cleaned and prepared data. Represent your findings visually as part of your submission.  
     
   I started by putting together a simple correlation matrix so that I could start picking out possible covariance between data elements, then attempted a Multivariate Analysis of Variance. MANOVA is a particularly appropriate choice in this case because I had not yet narrowed down the scope of my chosen dependent variables or independent variables and wanted to examine simultaneous effects.  
     
     
     
     
   With resp 1 representing Monthly Charges, it started to seem likely that I would want to focus on it as a main dependent variable since it appeared to be the most dramatically influenced of the three.   
     
   The above pieces of analysis also first flagged Senior Citizen to me as being the most impactful of the demographic variables I had decided to measure the effects of.   
     
   After looking at the MANOVA results, I still wanted to get another perspective of how these variables were “moving” together, so I put together a Principle Component analysis as well. I am a visual learner and the way that PCA is represented multi-dimensionally has, I have found, often helped me get a grasp of how to continue trying to group and explore any data I am working with.  
     
     
     
   PCA again confirmed for me that Monthly Charges was probably the most important dependent variable for me to focus on, and confirmed that it and Senior Citizen have a consistent relationship – in fact, SeniorCitizen may have the most consistent relationship with all three of the possible dependent variables I had decided to examine.  
     
   It is also interesting to note the possible relationships between MonthlyCharges and Paperless Billing, and between having Dependents and having a Partner. If I had more time I might look a little at how these things relate, especially in the context of the fuller dataset, however since Partner and Dependents are both variables I had intended to use as independents and MonthlyCharges and PaperlessBilling are both proposed dependent variables, I did not take it further.  
     
   It was at this point in the analysis that the pointlessness of attempting to focus on gender as an independent variable really revealed itself to me. Consistently between methods, it shows less correlation with any of the three dependent variables I wanted to examine than did Dependents and Partner and especially SeniorCitizen.  
     
   MonthlyCharges is, again consistently, the most readily influenced of the three dependent variables by far, so for the remainder of the analysis (with the exception of the multiple multivariate model I used to examine coefficients one last time) I focused on it alone. Though I continued to look at Partner and Dependents since they do show some apparent relationship to MonthlyCharges, I began to look more specifically at the relationship between a customer being a Senior Citizen and their monthly bill amount.   
     
   Lastly, using one of the regression models I had decided to use for prediction, I plotted the standardized residuals for a combination of the demographic variables I wanted to use as independents:  
     
     
     
     
     
   The Normal Q-Q plot shows a distinct bimodal distribution rather than a normal one – this is not surprising, as all of the independent variables the model was built to use are binary.   
     
   In some cases binary variables can compromise the accuracy of this type of regression model, however I have had good experiences using regression with binary variables regardless in similar circumstances, so rather than deviate from my initial plan I decided to evaluate the accuracy of the models after the fact using K-Fold Cross Validation instead, as you will see below.
2. Apply an analytic method and an evaluative method. Annotate the data showing both methods and your findings.  
     
   I initially put together a multivariate multiple regression by using cbind to use MonthlyCharges, Churn, and PaperlessBilling together. Below you can see the summary of that model.  
     
     
     
     
     
   The effect of being a Senior seemed to far outweigh the effects of having a Partner or Dependents when each effect was measured while controlling for the others (as in the model above). I ran a Wald test for both of those variables on a version of the model trained to a randomly sampled subset of the data, in order to evaluate their importance to the model. I already knew gender could almost certainly be removed safely thanks to my MANOVA and PCA results so I did not test it in the same way.

  
  
  
Somewhat surprisingly (to me) the P values for both were well below .05, causing me to reject the Wald test’s null hypothesis. This indicates that despite SeniorCitizen being the most important variable, it is at its predictive best used in combination with the other two.  
  
Still, I wanted to look at the impact of SeniorCitizen on only Monthly Charges taken in separate combination with each of the other two predictors, so I generated and plotted two more models plotting the effect of being a Senior combined with the effect of first being Partnered, and then with the effect of having Dependents.  
  
  


  
  
Having a partner and being a Senior have a reinforcing effect on one another, whereas, because having dependents is correlated with lower monthly charges, when combined with the SeniorCitizen variable it dampens the correlation between being Senior and having higher monthly bills.  
  
To give the clearest view of all, I looked at the effects of the three independent variables (SeniorCitizen, Partner, Dependents) on Monthly Charges separately, side by side:  
  
  
  
  
Finally, in order to address the potential issues with inaccuracy generated by my unconventional choice of using a linear model with binary independent variables, I decided to evaluate a MonthlyCharges ~ SeniorCitizen linear model using a form of K-Fold Cross Validation.  
  
  
   
  
  
  
As you can see, the means and predicted effects of SeniorCitizen on Monthly Charges within the cross validation results is very similar to what it has been for the applied linear models themselves. This confirmed my hunch that, given the simplicity of the data, using binary predictor variables would not significantly compromise the accuracy of linear or multivariate regression models.

1. Justify the methods you have chosen to analyze your data. Be sure to include details about how the methods you have chosen better represents your findings than other methods.  
     
   I chose to use MANOVA and PCA in the more exploratory/descriptive portion of my analysis because although I narrowed down my variables of interest early I retained a desire to look at more than a single dependent and independent variable at once. Also, I was unsure what relationships I would find within and between them initially. Both of these methods have a certain flexibility to them in terms of their ability to handle many data dimensions at once: PCA in particular was appealing because it did not require me to pre-define an amount of clusters or anything similar. I wanted the data itself to indicate to me its most natural groupings.  
     
   I chose to use combinations of Multivariate Multiple Regression and Linear regression for predictive analysis to look at combined effects of multiple independent variables on multiple dependent variables, and later to clearly represent the strongest linear relationships that I found. I like the interpretability and the high latency of this method and use it wherever I think it may be suitable for making basic predictions that I want to be easily understandable to viewers.   
     
   Although my use of binary variables (and the bimodal distribution it caused) might have complicated use of linear regression on some datasets, I showed that it did not significantly inhibit the accuracy of predictions on this data using K-Folds Cross Validation as an evaluative method.  
     
   I used the Wald test when I was unsure of whether to continue including two variables in my models, because it gave me an easy way to estimate how if at all the model would be impacted by their removal.
2. Justify the methods you have chosen to visually present your data. Be sure to include details about how the presentation methods you chose better represents your findings than other presentation methods.  
     
   All of the visualization methods I used were chosen for simplicity and clarity, essentially their readability.   
     
    For the univariate distribution representations I wanted clean, side by side representations of the data that would help me eyeball the distribution before testing it using fitDistr.   
     
   From the PCA, I chose the cleanest, most minimalistic means of representing the dimensionality of the data when multiple variables were accounted for simultaneously.  
     
   I used several correlation matrixes, from a basic one, to the one derived from the MANOVA, to the one derived from the PCA, because I think laying the numbers out in such an easily readable way helps viewers to interpret more abstract graphical representations as seen in the PCA chart.  
     
   I chose to include the Normal Q-Q plot because it demonstrated how including binary independent variables resulted in a non-normal distribution. I knew including it would engender explanation of my decision to use regression for prediction while also using binary variables, which needed addressing.  
     
   And I used simple regression line plots to visualize my regression models and my cross validation, because I think they’re the easiest way to communicate predictive trends. I used multiple visualizations for multiple models featuring different variables and combinations of variables so that effects could be visually compared.

***IV: Data Summary***  
Summarize the findings of your data evaluation. Provide the final findings dataset, including evaluation measures.  
  
The primary notable finding of my analysis is that being a Senior Citizen is associated with higher monthly bills. Visualizations of the findings and details on how I evaluated the models I developed are outlined in sections above.

1. Explain how your data shows that it was discriminating or not and whether the phenomenon you wanted to detect was present in your findings. Provide specific examples from the data to support your claims.  
     
   The results of the PCA analysis, seen again below, shows obvious patterns of discrimination between certain groups of variables.   
     
     
     
   The initial phenomenon I had thought to find in the data was a correlation between gender and MonthlyCharges, PaperlessBilling, or Churn. As also represented in the above graphic, that finding was not present. However I was correct that there would be some relationship between some of the demographic variables and MonthlyCharges – specifically I found a positive relationship between being a Senior Citizen and the amount spent per month.
2. Describe the methods you used for detecting interactions and for selecting the most important predictor variables. Include the specific interactions you detected and the most important predictor variables that you found.  
     
   I used PCA and MANOVA, as well as a simple correlation matrix, and later a bit of trial and error with models and Wald testing, to detect and evaluate the significance of various interactions. The most important predictor variable I found for any dependent variable was SeniorCitizen, although Dependents and Partner also showed some noticeable relationship to MonthlyCharges.
3. Acknowledge sources, using in-text citations and references, for content that is quoted, paraphrased, or summarized.  
     
   I made extensive use of the following R packages/libraries in this project:  
     
   library(FactoMineR)

library(factoextra)

library(mda)

library(MASS)

library(sjPlot)

library(rmutil)

library(effects)

library(lmtest)

library(survey)

library(caret)

library(e1071)

library(openxlsx)