Final Lab

Please use the headers to bounce around as needed

# Data Exploration

I started this project in SQL exploring the three csv files provided for the final project.

## A screenshot of a computer Description automatically generated SQL exploration and joining of CSV files

I attempted to join them together in SQL with the following code.

A screenshot of a computer

Description automatically generated

Which gave me the following error message and hint:

## A white background with red text Description automatically generatedA white rectangular sign with black text Description automatically generatedError message

I decided it would just be easier to comment out the top two lines and then save the following table as a csv file using the save results function.

## A screenshot of a computer Description automatically generatedSaving join CSV

Here are the top ten rows of my results.

A screenshot of a computer

Description automatically generated

I had to redo this part as I’m adding the visit date column because I didn’t include it before by mistake.

I explored the data a little bit to see if there were any obvious anomalies in the data set.

## A screenshot of a computer code Description automatically generatedFurther exploration

The major thing that I really noticed was there were only 98 patients in total out of the 500 visits. Patients 79 and 88 were missing from the data set, implying they had no appointments in 2023.

## A screenshot of a computer Description automatically generatedA screenshot of a number Description automatically generatedFound missing patient ID’s

I checked for NULL values and didn’t find any here’s an example of my code.

## A close up of a text Description automatically generatedChecked for NULL values

A screenshot of a computer

Description automatically generated

I moved over to RStudio after this, because I feel more familiar with it and wanted to get some more insight into what the data was telling me.

# Statistical Analysis RStudio

I imported my data and did a quick look at the data before manipulating it before and analysis.

### Importing joined CSV and setting up RStudio

A computer screen shot of a computer code

Description automatically generated

I also looked at the count of patients seen by age.

### Exploration

### A white text with black text Description automatically generatedA screenshot of a computer Description automatically generatedA screenshot of a table Description automatically generatedAge

It was interesting to see the spread of the population.

I conducted more counts to see which specialties saw more foot traffic.

### A screenshot of a computer Description automatically generatedSpecialization Counts

And which clinic saw the most foot traffic.

### A screenshot of a calendar Description automatically generatedClinicID Counts

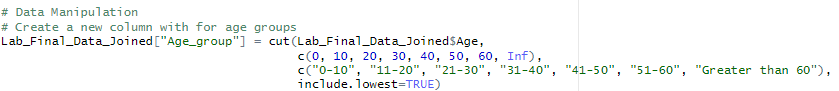
I also ran a test to see which patients were seen the most and the least.

### A screenshot of a computer program Description automatically generatedPatient Visit Count

After exploring the data, I added some columns to help with the string values so I could use them in my analysis.

## Data Manipulation

### Age Group Creation

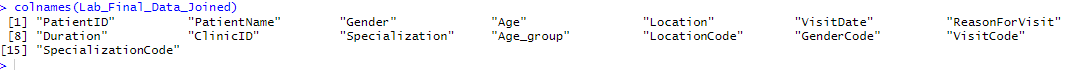


### For Loops to Add Numerical Representations of Character Values

A screenshot of a computer code

Description automatically generated

I did this for loop four times to create new columns for the reason for visit, location, gender, and the specialization visited by the patient.



Summarizing my data now showed the following.

### Data Summary

A close-up of a white background

Description automatically generated

I calculated the min, mean, max, standard deviation and variance of the duration spent at the clinic by each age group.

### Duration min, mean, max, SD, and Var

A screenshot of a computer code

Description automatically generatedA screenshot of a computer

Description automatically generated

### Duration by Specialty

A screenshot of a computer program

Description automatically generated

### Duration by Reason for Visit

A computer screen shot of a computer code

Description automatically generated

Because my new columns were chr values I had to convert them to numerical values. I did this after putting them into a new data frame I would use to do my correlation visualization.

### Correlation Test Set-upA screenshot of a computer code Description automatically generatedA black and grey text Description automatically generated

## Graphs

### Correlation plot

A graph of a number of codes

Description automatically generated

Which showed very little to no correlation between any of the compared values.

Next, I looked at some boxplots, histograms, and a density plot to visualize the data.



### Histogram

A graph of a number of bars

Description automatically generated

This graph visualizes the number of visits based on the patients age.



### BoxplotA graph of different colored lines Description automatically generated with medium confidence

This graph shows the duration spent at each location’s clinics by different age groups of patients. The dot indicates the mean duration for that specific location and age group.

A close up of text

Description automatically generated

### Density plotA graph of a graph Description automatically generated

The graph shows the density line curve by patient age with a dashed line representing the mean age of the patient population. Below shows which gender “dominates” each age group.

A white background with black text

Description automatically generated

A computer screen shot of a computer code

Description automatically generated

### Bar plot 1A screenshot of a graph Description automatically generated

This graph shows the total visits to specific specialties by each patient age group.

### Bar plot 2A graph of different colored squares Description automatically generated

This graph shows the total visits per month and which specialties saw the most patient traffic.

I tested the observed average duration of a patient’s visit by location to see if there were any significant findings.

## Hypothesis Testing

### t.test

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generatedA screenshot of a computer code

Description automatically generated

All the p-values heavily indicated that there is no statistical difference between the observed mean and the t-test calculations.

I tested by Age next.

A screenshot of a computer program

Description automatically generated

A screenshot of a computer code

Description automatically generated

A screenshot of a computer

Description automatically generated

The result was similar suggesting no statistical significance.

Next, I tested some two-sided tests based on location.

### Two-sided tests

A screenshot of a computer program

Description automatically generated

A screenshot of a computer program

Description automatically generatedA screenshot of a computer code

Description automatically generated

Which produced a not statistically significant result meaning the location seems to have little effect on the duration of a patient’s stay. Confirming the previously examined correlation graph.

### One-sided tests

Finally, I conducted two one-sided tests to see if gender had any effect on durations of appointments and age of the patient seen.

A computer screen shot of a code

Description automatically generated

A screenshot of a computer program

Description automatically generated

Both were statistically insignificant and thus had no influence on patients’ appointment duration or their age.

# Predictive Modeling

Now I started in Python working on building out my predictive model.

## Import data and packagesA screen shot of a computer program Description automatically generated

I re-explored my data in Python to make sure I didn’t miss any null values and get an idea of what values I was working with.

## Data exploration

A screenshot of a computer program

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generatedA screenshot of a computer screen

Description automatically generatedA computer screen shot of a computer code

Description automatically generated

## Modeling Prep

I created dummy variables for my modeling process.A screenshot of a computer

Description automatically generated

This is where I’m still confused on how to properly test each of my variables of interest. I honestly want to look at duration, reason for visit, location, and specialty seen, but I couldn’t figure out how to write a program that works with a non-binary y value. I kept getting an Error telling me to make it a multiclass of either ovo, or ovr, which I just couldn’t figure out how to execute that with the information I’ve found.A screenshot of a computer program

Description automatically generated

## Running models

A screenshot of a computer program

Description automatically generated

My decision tree model seemed to produce the most accurate result for predicting gender based off the X values provided as his was the only y value I could get to work with the code I wrote.

# Actionable Insights

## Insights

* The average age of patients seen was 43 years old. Male average age is 41 and the female average age is 46 years of age.
* Average time spent at any appointment by a patient was 93.58 minutes. Males on average waited 94.19 minutes and females waited on average 92.69 minutes.
* Clinics 4, 2, and 8 saw the most patient foot traffic
* The most visited clinics were Neurology and Pediatrics
* Patients 79 and 88 are missing from this year’s recordings, possibly an error in identifying patients upon intake, but more likely these patients were not seen this whole year as there was no miss aligned data indicating an inaccurate recording.
* The age group with that largest standard deviation (52.28) and variance (2733.05) of the duration of their appointments is 31- to 40-year-olds.
* The density plot shows that the distribution of patients by age and gender has younger male patients and more elderly female patients.

## Recommendations

* Patient ID’s 79 and 88 could just be a recording error or patients that failed to attend an appointment all year. Regardless of which issue it is, the hospital/clinics should implement a quality assurance program that would help eliminate possible clerical errors through a recording method that indicates whether a patient was “on-time”, “late”, or “no-show”. This would help to record usable data for future analysis.
* The average time spent at the hospital was just over an hour and a half for both females and males. This holds true for each specialty and for the reason the patients visited the hospital. Depending on the hospital’s goal for appointment scheduling a couple of different suggestions can be made. First, a more accurate recording of time should be implemented, “check-in time”, “time until called back to an exam room”, “in exam room before doctor”, “in exam room with doctor”, “check-out time”. This breakdown would give a better reading of where and how improvements can be made, or better time management can be implemented. This way patient wait times can be minimized. The long appointments could also be due to short staffing, which a more precise documentation of the patient’s pathway through the hospital might help identify.
* The patient population seen at this hospital suggests that there are minimal young adult female patients compared to male. A lack of women’s health clinical services such as Gynecology or labor and delivery would indicate a potential area of growth for the hospital to increase it’s patient population.