Simran Bhalla

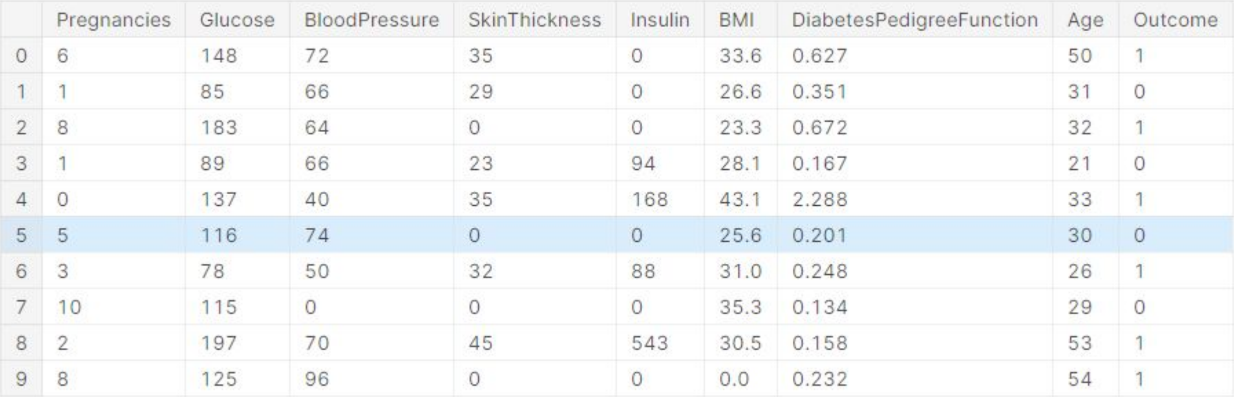
November 5th 2020

Data Mining - Guha

Data Exploration and Visualization

When one begins to analyze data, it is important to explore your data before going forth and implementing complex functions to the data. Descriptive statistics is one way to do this. With descriptive statistics, we describe what the best features of the data are. We also look into the samples and measures of the data whenever possible. We want to be able to tell what the distribution between different variables are and what that means for certain values like the standard deviation, mean, max, and min. Exploratory data analysis allows the user to use a set of procedures to refer to a set of data. It is useful because it gives descriptive and graphical summaries of the data. You can look at the data without making any previous assumptions.

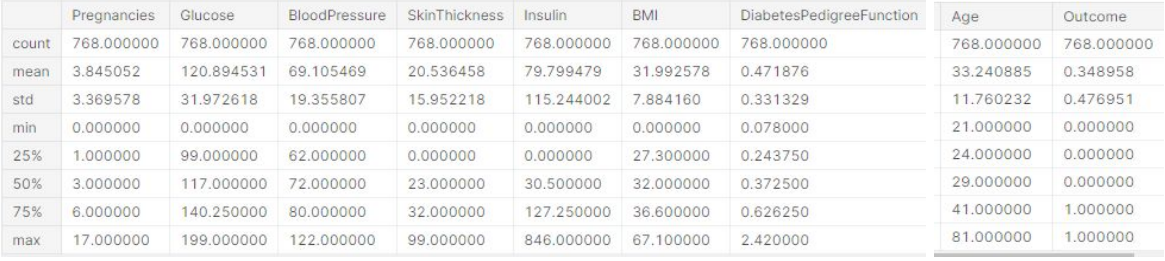
In this project I was able to use our extensive learning of data analysis and I used the question of whether the patient was likely to be diabetic or not as my focus. There were 7 categories that I had talked about analyzing and looking through in the last deliverable. There are pregnancies, glucose level, blood pressure, skin thickness, insulin level, BMI, and age. I selected 10 patients that were given and wanted to see what the predicted outcome would be at first. We talked in class about how it is important to look through all of the categories and how they will all lead to a distinct outcome. To see how the answers would change, I deleted some columns here and there to see if the outcome would change and it did. This was very interesting for me to see. Below is a snipping of what my data looked like after.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
0	6	148	72	35	0	33.6	0.627	50	1
1	1	85	66	29	0	26.6	0.351	31	0
2	8	183	64	0	0	23.3	0.672	32	1
3	1	89	66	23	94	28.1	0.167	21	0
4	0	137	40	35	168	43.1	2.288	33	1
5	5	116	74	0	0	25.6	0.201	30	0
6	3	78	50	32	88	31.0	0.248	26	1
7	10	115	0	0	0	35.3	0.134	29	0
8	2	197	70	45	543	30.5	0.158	53	1
9	8	125	96	0	0	0.0	0.232	54	1

Figure 1 data for each category with outcome proected

From here I decided to create the describe method. This helped me determine what the mean, standard deviation, min, max, and quartile values would be. I implemented this to try and figure it out for all of the different categories. This was really helpful to help me analyze the numeric and object series. A picture of the screenshot of what running this function would look like is below.



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
count	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000	768.000000
mean	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.471876	33.240885	0.348958
std	3.369578	31.972618	19.355807	15.952218	115.244002	7.884160	0.331329	11.760232	0.476951
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.078000	21.000000	0.000000
25%	1.000000	99.000000	62.000000	0.000000	0.000000	27.300000	0.243750	24.000000	0.000000
50%	3.000000	117.000000	72.000000	23.000000	30.500000	32.000000	0.372500	29.000000	0.000000
75%	6.000000	140.250000	80.000000	32.000000	127.250000	36.600000	0.626250	41.000000	1.000000
max	17.000000	199.000000	122.000000	99.000000	846.000000	67.100000	2.420000	81.000000	1.000000

Figure 2:
data of the describe method created

For this iteration I was most interested in the glucose levels for diabetes because based off of my data and what we know about diabetes I know that this is the most effective category. I have created a boxplot because I know that this is a great way to display the data that is shown. We are able to see what the outliers are and clearly see where the mean and median lie. My boxplot is shown below. Again the values are for glucose and the outcome.

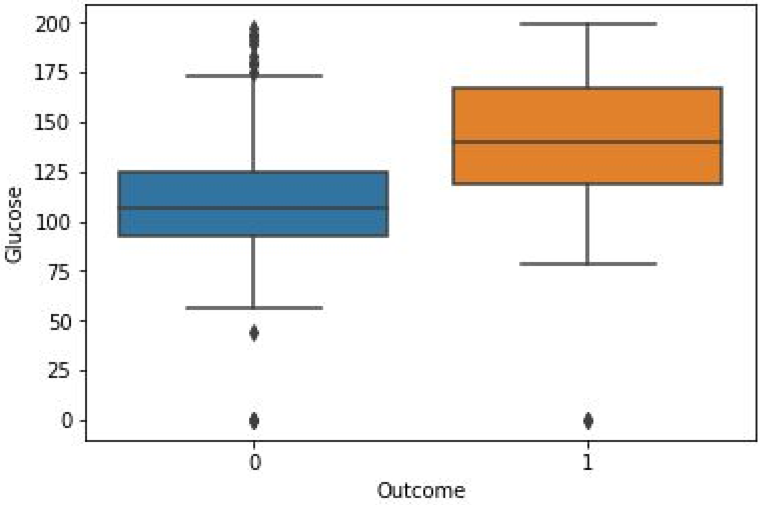


Figure 3: boxplot data created for glucose and outcome

This data was very valuable to me because I was able to see clearly how in between 115 and 125 the outcome was most likely to change from 0 to 1. I really am glad that I chose this as a key category to look at and it puts what we talked about in class into perspective. It also got me thinking about other prior readings and examples we have conducted and how seeing a representation like this could be incredibly valuable.

The next diagram and graph I was able to create was the histogram. I decided to use this one because it allows me to analyze the ones and the zero outcomes in more depth. I again decided to use the glucose category to compare it to the outcome. The histogram data that I used here shows frequency distribution values.

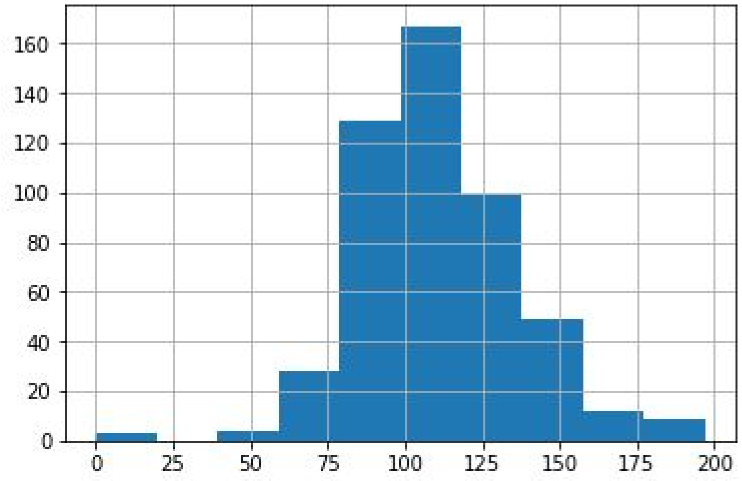


Figure 4: Histogram created for output = 0

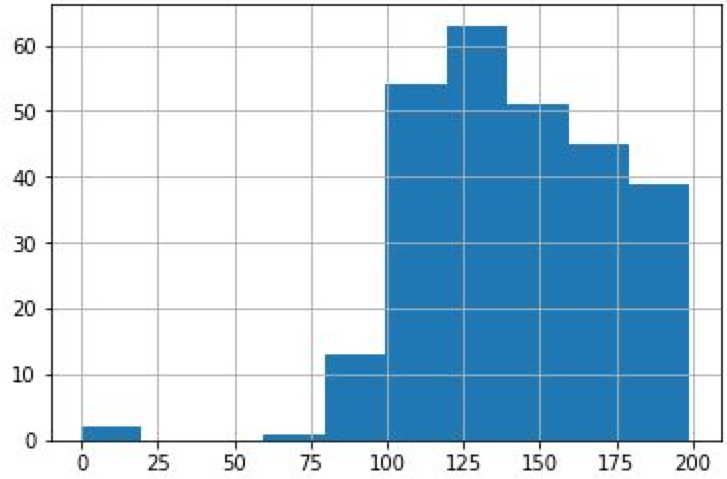


Figure 5: Histogram created for output = 1

The last data exploration tool I decided to use was the scatterplot. I created a scatter plot diagram of all of the glucose values and the BMI values on different axes and had the points be placed for the outcome value for and 1. Overall this was an interesting tool to look at. The image for it is placed below.

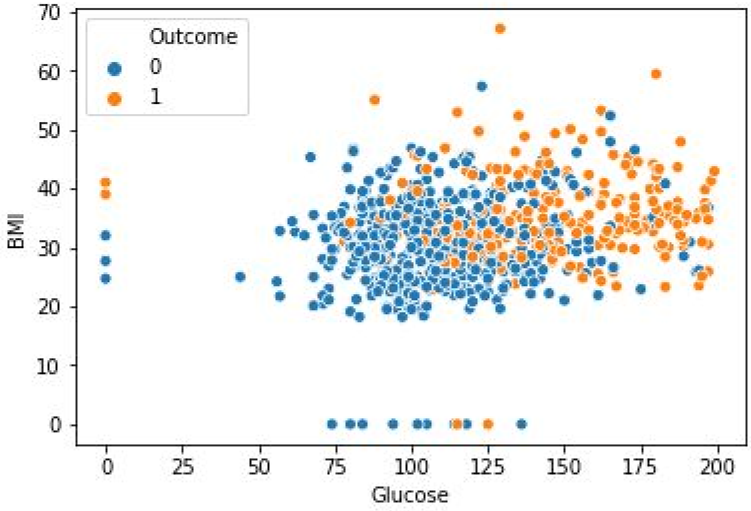


Figure 6: Scatterplot created to compare BMI and Glucose

If I had more time to continue this project I would continue to explore different categories and see what the data visualization outcomes could look like. Creating these helped reinforce everything that we have been learning and allowed me to get hands in experience. I couldn’t help but wonder what these data visualization outcomes would look like for projects like the AlphaGo project created by DeepMind. We had to watch the documentary for my Artificial Intelligence class and I couldn’t help but wonder what the data visualizations would look like for that. It is a fascinating concept that is prevalent all around us now a days

Sources

[1] Frost, J., Adusei, C., Peeyush, Siyabonga, Sreeja, Narayanan, J., . . . Sachin. (2019, June 13). Identifying the Most Important Independent Variables in Regression Models. Retrieved October 08, 2020, from https://statisticsbyjim.com/regression/identifying-important-independent-variables/

[2] T anyildizderya. (2019, September 10). Diabetes Prediction with Logistic Regression. Retrieved October 08, 2020, from https://www.kaggle.com/tanyildizderya/diabetes-prediction-with-logistic-regression

[3] What is Logistic Regression? (n.d.). Retrieved October 08, 2020, from https://www.statisticssolutions.com/what-is-logistic-regression/

Data visualization beginner's guide: A definition, examples, and learning resources. (n.d.). Retrieved November 06, 2020, from https://www.tableau.com/learn/articles/data-visualization