Christopher Dillard 2/10/2024

Understanding diabetes progression through an analytical lense: A comparative report on the applications of JMP, Tableau, and R for descriptive and predictive analytics:

Throughout this report, I will be sharing my interpretation of a multivariate analysis that explores the significance of several physiological markers in determining a patient's likelihood of developing high diabetic disease progression after one year of recording their baseline values. By employing predictive models and visualization techniques, my research serves the purpose of extracting meaningful insights into the dynamics of diabetes progression while primarily showcasing the capability of JMP as a powerful tool for building predictive models. This research does not constitute medical advice, but rather focuses on how clinical data can be analyzed using the knowledge that I have acquired throughout a master's degree in business Analytics. Along with JMP, I will be delving into a comprehensive analysis of the capabilities of Tableau and, to a lesser extent, RStudio to demonstrate that employing a multi-software approach is the optimal method for analyzing these data. This will involve comparing and contrasting aspects such as output interpretability, visual appeal, and each software's inherent capabilities for specific tasks. To this end, I will provide alternative versions of some supporting visualizations and tables created using these programs with brief descriptions highlighting instances where one software outperforms another in its analytical capabilities for the specific purpose, thus advocating for a more diversified analytical approach.

Throughout this report, I will be referencing a study conducted with a cohort of 442 patients with diabetes (refer to the second reference on page 23 and follow the link to download the full journal article). This dataset contains recorded levels of 10 baseline variables, with a mix of qualitative and quantitative properties that were recorded and assigned to each patient, as a means to assess how these values impact the progression of the patient's disease after one year. Initially, I will explain any analysis conducted in JMP, where the focus is geared towards creating a predictive model with only the most relevant variables in determining disease progression, particularly in how their recorded values could contribute to the level of diabetic disease progression, which is stored in the response variable's column, "Y Binary" with a binomial classification of LOW, indicating relatively better patient outlook one year after baseline levels are recorded, and HIGH, representing higher disease progression. There are a total of 442 patients in the dataset, aged 19-79 (recorded within the "Age" column). The values in "Y Binary" are directly based off of another column in the dataset, "Y", which is described as a "Quantitative measure of disease progression one year after baseline [values were recorded]". Under the study's criteria, any Y values over 200 signifies High disease progression, while values less than or equal to 200 signify low levels of disease progression. Gender is encoded by 1 or 2, as information about which number corresponds to which gender is not disclosed. The gender split is about a 0.53 and 0.47 for genders "1" and "2", respectively; so, this is a fairly even split. Nondisclosure of such a variable could be helpful to discourage confirmation bias when running predictive models, in terms of how the results are interpreted. That being said, the absence of clear gender information may hinder deeper interpretability of the models in case there are any genuine medical implications as for whether gender has any statistical importance in expected ranges for values for any other variable, which we can infer by looking at how JMP selects its variables of importance in the final predictive model.

Another variable that is explored in this assignment is BMI (Body Mass Index). Any BMI between 25.0 to 29.9 is considered overweight, while anything above that range is considered obese (cdc.gov, 2022). In this dataset, the mean BMI is 26.4, meaning that the average BMI is well within the overweight range. Within JMP, I produced the following visualization using the Distribution icon after selecting the BMI column to explore the distribution of values within variables for all patients, regardless of their level of disease progression. Note here that JMP includes both a histogram and boxplot to create a visually simplistic, yet highly informative visualization of the distribution of data points for BMIs, where I easily changed the color of the histogram's bins to blue and added labels above each bin for the percentage of data points that fall within each bin's range:

Figure 1: Body Mass Index (BMI) Distribution

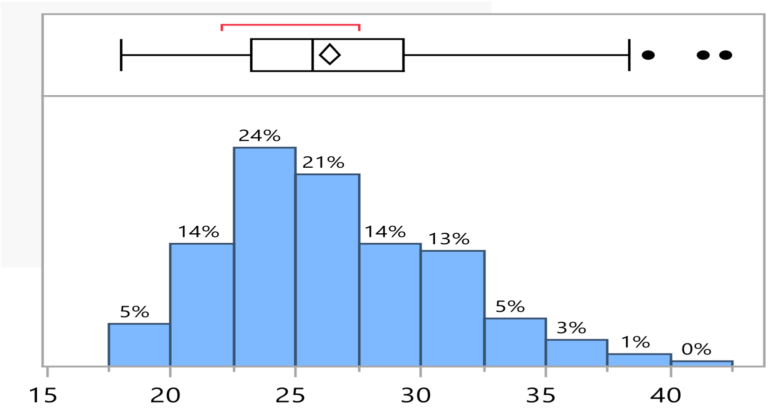
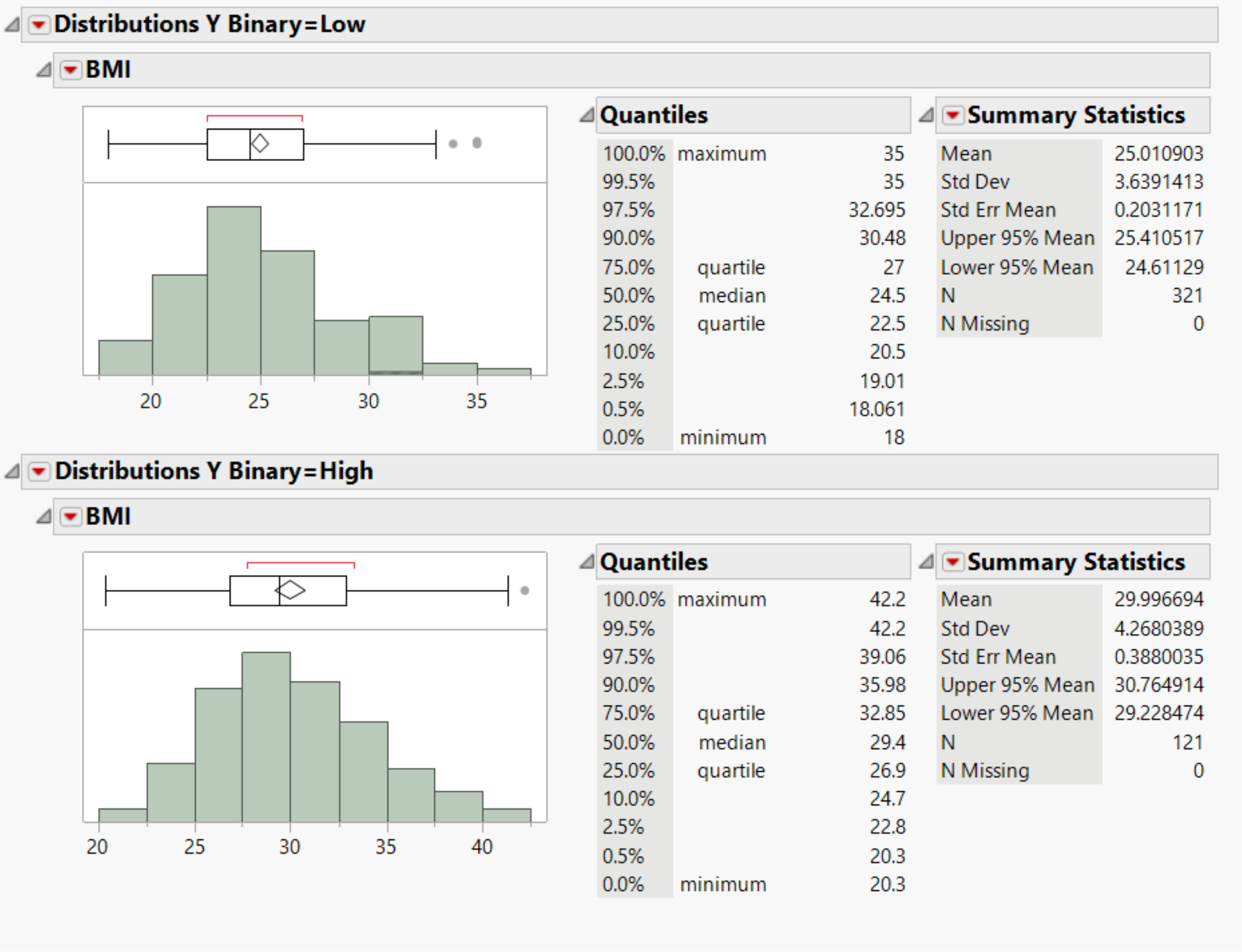


Figure 2: BMI Quantiles from JMP

100.0%	maximum	42.2
99.5%		40.827
97.5%		36.1
90.0%		32.37
75.0%	quartile	29.325
50.0%	median	25.7
25.0%	quartile	23.175
10.0%		21
2.5%		19.3075
0.5%		18.186
0.0%	minimum	18

The above histogram reveals that the distribution of BMIs is skewed to the right, meaning that the spread of these data is further to the right side of the median (i.e., the higher values drive the mean to be higher than the median). The data within the "Quantile" table tells us that 50% of the patients have BMIs above 25.7, and almost 25% are considered obese. With so many of the patients being overweight, in addition to the large spread of values above the mean, the model's selection of BMI and its estimation formula could give better insight into just how significant BMI may be as a predictor variable of diabetic disease progression. By producing two separate histograms for patients in the High and Low groups of disease progression, some clearer differences begin to arise:

Figure 3: BMI Distributions Split by Progression Levels



Distributions Y Binary=Low

BMI

Quantiles

100.0%	maximum	35
99.5%		35
97.5%		32.695
90.0%		30.48
75.0%	quartile	27
50.0%	median	24.5
25.0%	quartile	22.5
10.0%		20.5
2.5%		19.01
0.5%		18.061
0.0%	minimum	18

Summary Statistics

Mean	25.010903
Std Dev	3.6391413
Std Err Mean	0.2031171
Upper 95% Mean	25.410517
Lower 95% Mean	24.61129
N	321
N Missing	0

Distributions Y Binary=High

BMI

Quantiles

100.0%	maximum	42.2
99.5%		42.2
97.5%		39.06
90.0%		35.98
75.0%	quartile	32.85
50.0%	median	29.4
25.0%	quartile	26.9
10.0%		24.7
2.5%		22.8
0.5%		20.3
0.0%	minimum	20.3

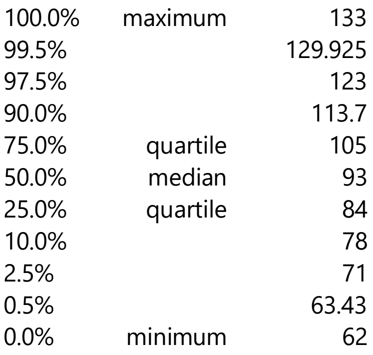
Summary Statistics

Mean	29.996694
Std Dev	4.2680389
Std Err Mean	0.3880035
Upper 95% Mean	30.764914
Lower 95% Mean	29.228474
N	121
N Missing	0

From the quantiles above, with again a minimum threshold of 25 BMI for being overweight and any BMI over 29 constituting obesity, it is clear that only 50% of the patients in the Low disease progression group were at least overweight and the upper 10% were obese. As for the "High" group, over 75% of the patients were at least overweight and nearly 50% were obese. This large discrepancy in values between the two groups suggests that BMI should be an important predictor variable for the model.

One variable that unfortunately does not appear to make any medical sense in how it was recorded in the dataset, and therefore possibly statistically unreliable by default without some degree of assumption on my part, is BP (Blood Pressure). Blood pressure is typically recorded as a "fraction" with the systolic blood pressure, the numerator in the fraction, and diastolic blood pressure, the denominator (Heart.org, 2023). Both numbers need to appear together, unless otherwise specified, since the systolic reading is the blood pressure when the heart is contracting and diastolic is the “resting” blood pressure. Furthermore, any division of systolic readings by their diastolic counterparts would not realistically produce the values seen in the BP column. Upon further examination of the data’s source (cited as “Efron, Bradley, et. al”., 2004), I could not find any clarification regarding the BP variable; so, I would presume that the values correspond to the diastolic blood pressure, again the denominator of a typical BP reading, since a diastolic reading of less than 80 is normal and over 120 is considered a hypertensive crisis, otherwise constituting a medical emergency. In the table below, we can see that only 2.5% of patients in the sample had values above 123 while the rest fell within a reasonably realistic range for diastolic BP readings. This interpretation is purely my speculation, so, if BP is selected as a predictor variable by the best model in JMP, I would limit my discussion of any finding of significance in terms of BP’s relevance as a predictor variable, rather than what the individual values mean.

Figure 4: Blood Pressure (BP) Quantiles



100.0%	maximum	133
99.5%		129.925
97.5%		123
90.0%		113.7
75.0%	quartile	105
50.0%	median	93
25.0%	quartile	84
10.0%		78
2.5%		71
0.5%		63.43
0.0%	minimum	62

The other 6 baseline variables include Total Cholesterol, LDL (Low Density Lipoproteins), HDL (High-Density Lipoproteins), TCH (Total Cholesterol divided by HDL), LTG (Logarithm of Triglyceride level), and Glucose. Should these variables be selected by the model, I will interpret their significance after revealing their corresponding values in the best predictive model’s equation, along with their signage, which would reveal the direction of their relation to the response variable (i.e. whether an increase in the values tend to create an increase in the likelihood of high diabetic disease progression or not). “Best” is used here to denote that the selected predictive model had the most accurate predictive capabilities on data split that it was not trained on within JMP.

Classification problems look for the probability of different outcomes that are not continuous (not time-series data) and results are interpreted as the probability of certain outcomes, which in this case would only be either LOW or HIGH disease progression. Given that my interest is in seeing how the various variables could predict whether a patient's diabetic condition progresses (gets worse), I will be focusing on the prediction formulas that correspond to HIGH predictions after a year for all patients and disregard the predicted LOWs.

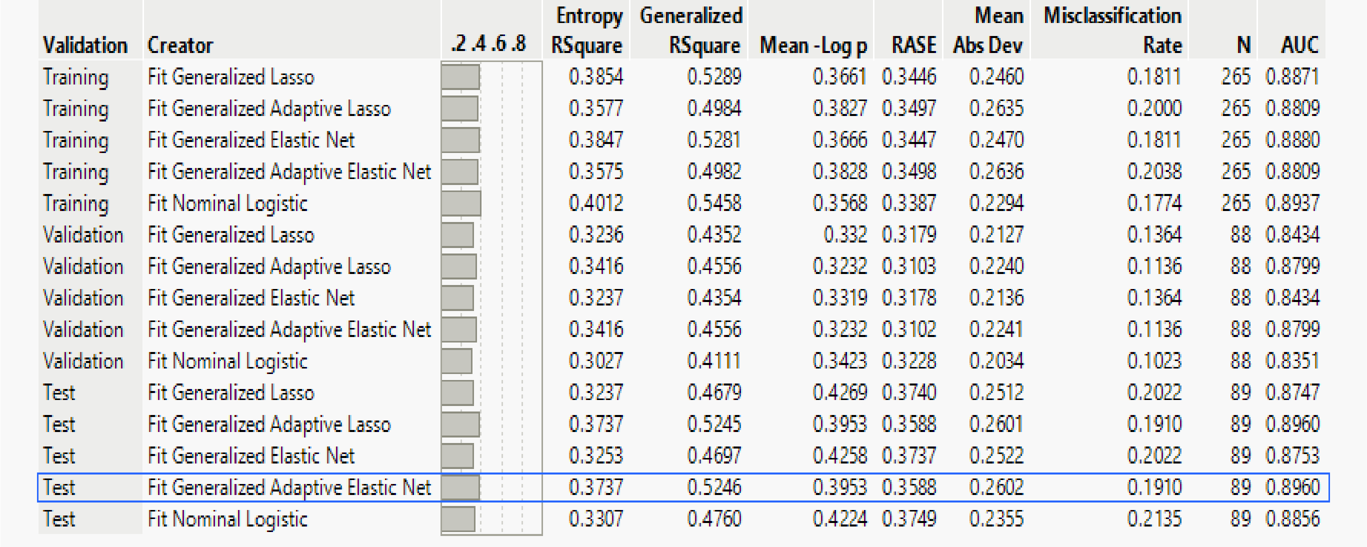
I will employ 5 different methods in this assignment, one being Ordinary Logistic Regression and the rest belong to the Penalized Regression class of models: Lasso, Adaptive Lasso, Elastic Net, and Adaptive Elastic Net. One factor that must be addressed when selecting the best model, is how the ratio of rows to variables, 442 to 10, leaves me with the concern that there is a potential for overfitting. So, variable selection is of utmost importance when there is a chance of this phenomenon. Previous assignments that I completed have proven that Lasso and Elastic Net models and their adaptive versions do very well with variable selection by eliminating less important variables, something that the Ordinary Logistic Regression model would not be as efficient at doing.

Given the cross-sectional nature of the data (not a time-series sample), I will be implementing a random seed split of 123 on patients' data to ensure model reproducibility, using a 60-20-20 split for the training, validation, and testing sets. A new column called "Validation" will be used to easily identify each patient by their assigned split.

The Area Under the Curve (AUC) will be my primary focus when analyzing the output for the Modeling Comparison tool in JMP. AUCs explain how efficient classification models are at predicting correct outcomes on a range between 0 and 1, where 1 would indicate that the model is a perfect fit. Any model above 0.5 can be interpreted as being more effective at predicting outcomes than a coin flip (i.e., random chance), and anything 0.5 and below is automatically useless for the sake of this assignment.

The image below captures the model comparison output, where my focus is on the AUCs for each model on the test set.

Figure 5: Model Comparison Tool in JMP

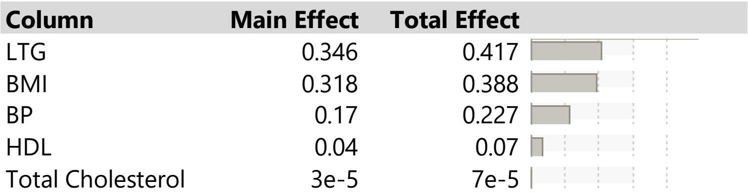


Validation	Creator	.2 .4 .6 .8	Entropy RSquare	Generalized RSquare	Mean -Log p	RASE	Mean Abs Dev	Misclassification Rate	N	AUC
Training	Fit Generalized Lasso		0.3854	0.5289	0.3661	0.3446	0.2460	0.1811	265	0.8871
Training	Fit Generalized Adaptive Lasso		0.3577	0.4984	0.3827	0.3497	0.2635	0.2000	265	0.8809
Training	Fit Generalized Elastic Net		0.3847	0.5281	0.3666	0.3447	0.2470	0.1811	265	0.8880
Training	Fit Generalized Adaptive Elastic Net		0.3575	0.4982	0.3828	0.3498	0.2636	0.2038	265	0.8809
Training	Fit Nominal Logistic		0.4012	0.5458	0.3568	0.3387	0.2294	0.1774	265	0.8937
Validation	Fit Generalized Lasso		0.3236	0.4352	0.332	0.3179	0.2127	0.1364	88	0.8434
Validation	Fit Generalized Adaptive Lasso		0.3416	0.4556	0.3232	0.3103	0.2240	0.1136	88	0.8799
Validation	Fit Generalized Elastic Net		0.3237	0.4354	0.3319	0.3178	0.2136	0.1364	88	0.8434
Validation	Fit Generalized Adaptive Elastic Net		0.3416	0.4556	0.3232	0.3102	0.2241	0.1136	88	0.8799
Validation	Fit Nominal Logistic		0.3027	0.4111	0.3423	0.3228	0.2034	0.1023	88	0.8351
Test	Fit Generalized Lasso		0.3237	0.4679	0.4269	0.3740	0.2512	0.2022	89	0.8747
Test	Fit Generalized Adaptive Lasso		0.3737	0.5245	0.3953	0.3588	0.2601	0.1910	89	0.8960
Test	Fit Generalized Elastic Net		0.3253	0.4697	0.4258	0.3737	0.2522	0.2022	89	0.8753
Test	Fit Generalized Adaptive Elastic Net		0.3737	0.5246	0.3953	0.3588	0.2602	0.1910	89	0.8960
Test	Fit Nominal Logistic		0.3307	0.4760	0.4224	0.3749	0.2355	0.2135	89	0.8856

As seen in the table, the AUCs are identical for both the Adaptive Lasso model and the Adaptive Elastic Model on the Testing set, which is of course the "new" and unbiased 20% of the data that the model was not trained on. This constitutes the need for further analysis of both models to see which one would be the best fit for the data. The only value that would favor one of these models over the other in this case has been shown by the blue box in the table: the generalized R Square for the Adaptive Elastic Net is 0.0001 higher than the Adaptive Lasso's value. So, I opted to choose the Adaptive Elastic Net as the best model for this case. It is also worth noting that the performance metrics for all the models do not appear to have a large degree of variation across the board; but the Ordinary Logistic Regression (Nominal Logistic) did perform the worst out of the 5 models in its predictive capabilities. By reviewing the AUC, we can see that the Adaptive Elastic Net model was able to accurately predict nearly 90% of the patient outcomes in the testing split, or 88 patients (20% of the entire sample).

After making this decision, I then explored each variable's importance in the model to see where analysts should primarily direct their attention to when discussing patient outlook and diabetic disease progression.

Figure 6: Adaptive Elastic Net Variables of Importance

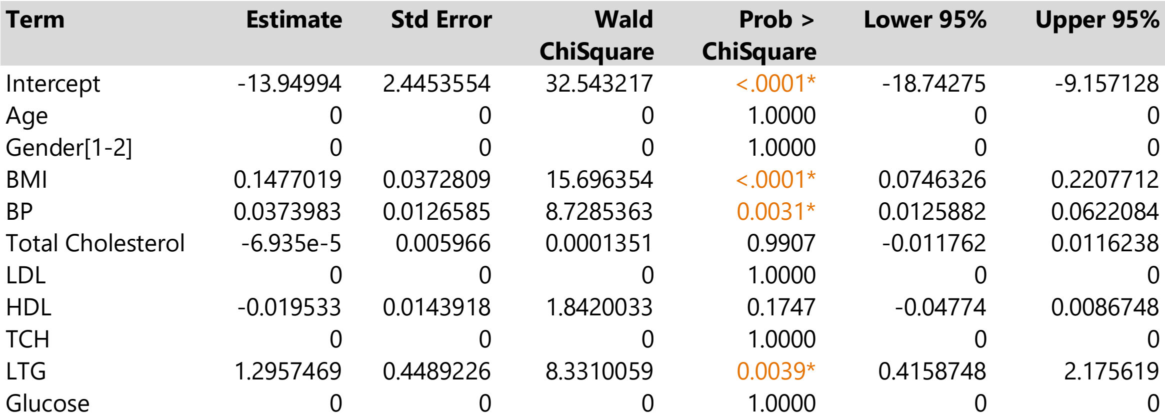


Column	Main Effect	Total Effect
LTG	0.346	0.417
BMI	0.318	0.388
BP	0.17	0.227
HDL	0.04	0.07
Total Cholesterol	3e-5	7e-5

The Adaptive Elastic Net model identified, in descending order, LTG, BMI, BP, HDL, and, to a very small degree, Total Cholesterol as the most important variables out of the ten baseline variables in the dataset, essentially deeming the rest to be statistically irrelevant in predicting disease progression compared to the 5 variables in the above table.

As shown by the parameter estimates, Age, Gender, LDL, TCH, and Glucose were dropped from the model. With gender being dropped by the model, this makes my previous concern about interpretability without the knowledge of which number corresponds to which gender irrelevant.

Figure 7: Parameter Estimates for Original Predictors - Adaptive Elastic Net Model



Term	Estimate	Std Error	Wald ChiSquare	Prob > ChiSquare	Lower 95%	Upper 95%
Intercept	-13.94994	2.4453554	32.543217	<.0001*	-18.74275	-9.157128
Age	0	0	0	1.0000	0	0
Gender[1-2]	0	0	0	1.0000	0	0
BMI	0.1477019	0.0372809	15.696354	<.0001*	0.0746326	0.2207712
BP	0.0373983	0.0126585	8.7285363	0.0031*	0.0125882	0.0622084
Total Cholesterol	-6.935e-5	0.005966	0.0001351	0.9907	-0.011762	0.0116238
LDL	0	0	0	1.0000	0	0
HDL	-0.019533	0.0143918	1.8420033	0.1747	-0.04774	0.0086748
TCH	0	0	0	1.0000	0	0
LTG	1.2957469	0.4489226	8.3310059	0.0039*	0.4158748	2.175619
Glucose	0	0	0	1.0000	0	0

As for the most important predictor variable, LTG, or the logarithm of triglyceride level, high levels of triglycerides in the blood are often indicative of metabolic disorders (National Heart Lung and Blood Institute, 2023). This implication is significant since this variable represenst a type of lipid (fat) that the body uses to store calories that it doesn’t immediately need after eating. Conditions like Diabetes directly impact how the body processes and stores energy and in patients with this disease, particularly Type 2 Diabetes, insulin resistance is a factor that leads the body to resort to other sources of energy when there is some inability to correctly use insulin to bring glucose, the general primary source of energy, to the cells. In other words, the body releases triglycerides into the bloodstream as an alternative source of energy to glucose and higher levels of triglycerides can be indicative of higher resistance to insulin.

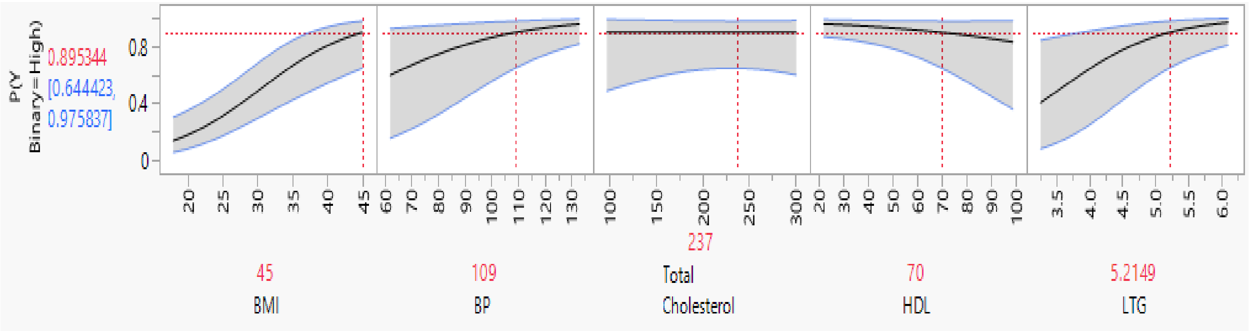
I was initially surprised to see that Glucose was not selected as a predictor variable since taking glucose levels is usually routinely done by diabetic patients to manage their disease. One potential reason for the exclusion of this variable in the model could have to do with how glucose levels are taken in the laboratory setting, along with how glucose levels fluctuate throughout the day, depending on what a patient is eating and when they're last meal was relative to when their bloodwork was taken (Nekrani et. al., 2023). Since no mention of timing is available in the dataset my interpretation would be purely speculatory; but, if lab work was done in the morning, patients with worse disease progression could have elevated numbers if their last meal was closer to when they fell asleep or whether they ate anything for breakfast prior to getting bloodwork done. So, a diabetic patient could reasonably be aware of this and already have adjusted their eating schedule accordingly which could give a lower glucose level. Furthermore, we could assume that some patients did not eat in the hours prior to their appointment, but otherwise are less conscientious about their eating habits, thus facing higher disease progression despite having a low blood glucose reading in a clinical setting. So, the exclusion of glucose from the model could be reflective of how blood glucose readings significantly fluctuate in any given day, which could render this variable to be too unreliable for a predictive model.

Now, I can use JMP to develop a hypothetical prognosis on a patient that has the following profile:

Age	Gender	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose
47	1	45	109	237	100.2	70	3	5.2149	107

JMP streamlines this process by allowing analysts to interpret the parameter estimates by viewing the prediction profilers of their model. Then, hypothetical values can be inputted to automatically view the probability of a certain outcome, being the HIGH predictions in this case. Note below that the values for Age, Gender, LDL, TCH, and Glucose are not factored in the model's predictions:

Figure 8: Prediction Profilers for Model Predictions for Hypothetical Patient



This result tells us that the patient in question would be at an 89.5% risk (probability) of advancing to higher disease progression in one year based on their baseline values. As I discussed previously from the parameter estimates table, some of the information on the patient was deemed to be irrelevant in the model and this is apparent where their values are substituted by a zero across the rows corresponding to the patient's age, gender, LDL, TCH, and Glucose levels.

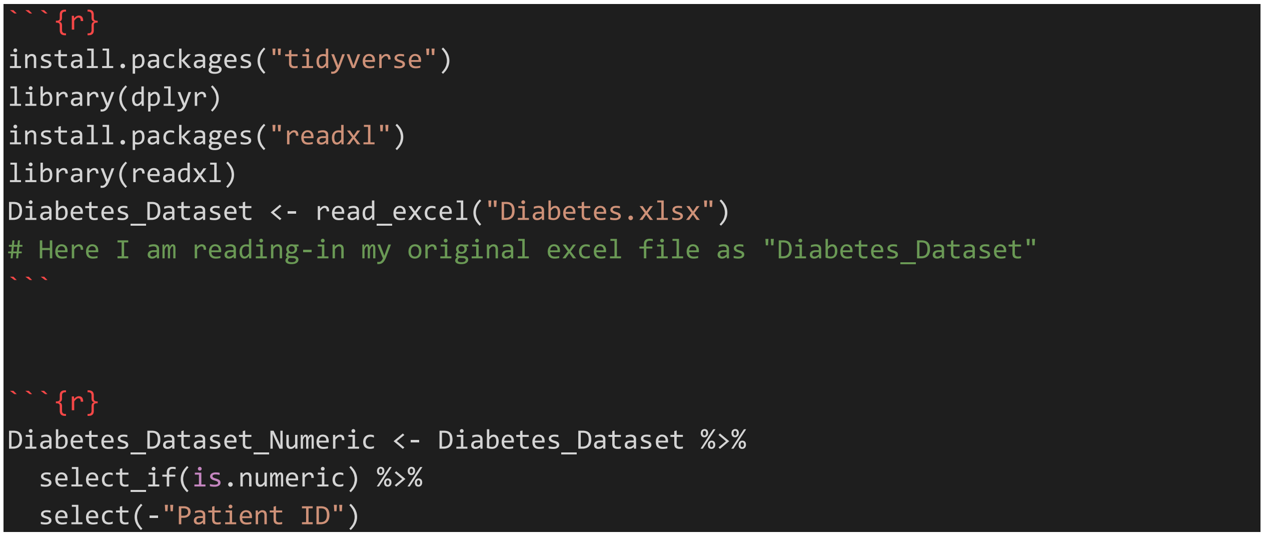
So far, I have explored the power of JMP as a predictive analytical software. Using the same data, I wanted to explore the relationships between the predictor variables from the adaptive elastic net model with each other, as well as with the response variable for disease progression in both its binary version and the numerical representation. With this in mind, I uploaded the dataset into RStudio to build two additional visualizations that better assess the importance of each variable in the model using two types of correlation matrices.

In simple terms, correlation matrices display the relationship among each variable, which could identify issues such as multicollinearity (i.e. when predictor variables are highly related to each other), which may hinder model's interpretability in terms of each variable's individual relationship with the response variable "Y". Although multicollinearity does not necessary lead to another potential issue, Overfitting, it could exacerbate this issue. Overfitting is essentially when a model's predictive capabilities are significantly poorer on the testing data split than on the training data. With an additional “validation” split (at 20% of the data) in the model’s building process, there is a buffer to mitigate the risk of overfitting by actively comparing the predictive capabilities of the model on new data before settling on an equation. Overfitting can arise when there are too many variables in the model and too complex of relationships between variables, or even not enough observations relative to the number of variables in the model. With 60% of the training data coming out to roughly 265 patients, this is far from constituting a large study, however the cohort does comprise a fairly large degree of variation for recorded values of each variable, as can be explored by histograms in JMP or, as I will show later in this paper, in Tableau.

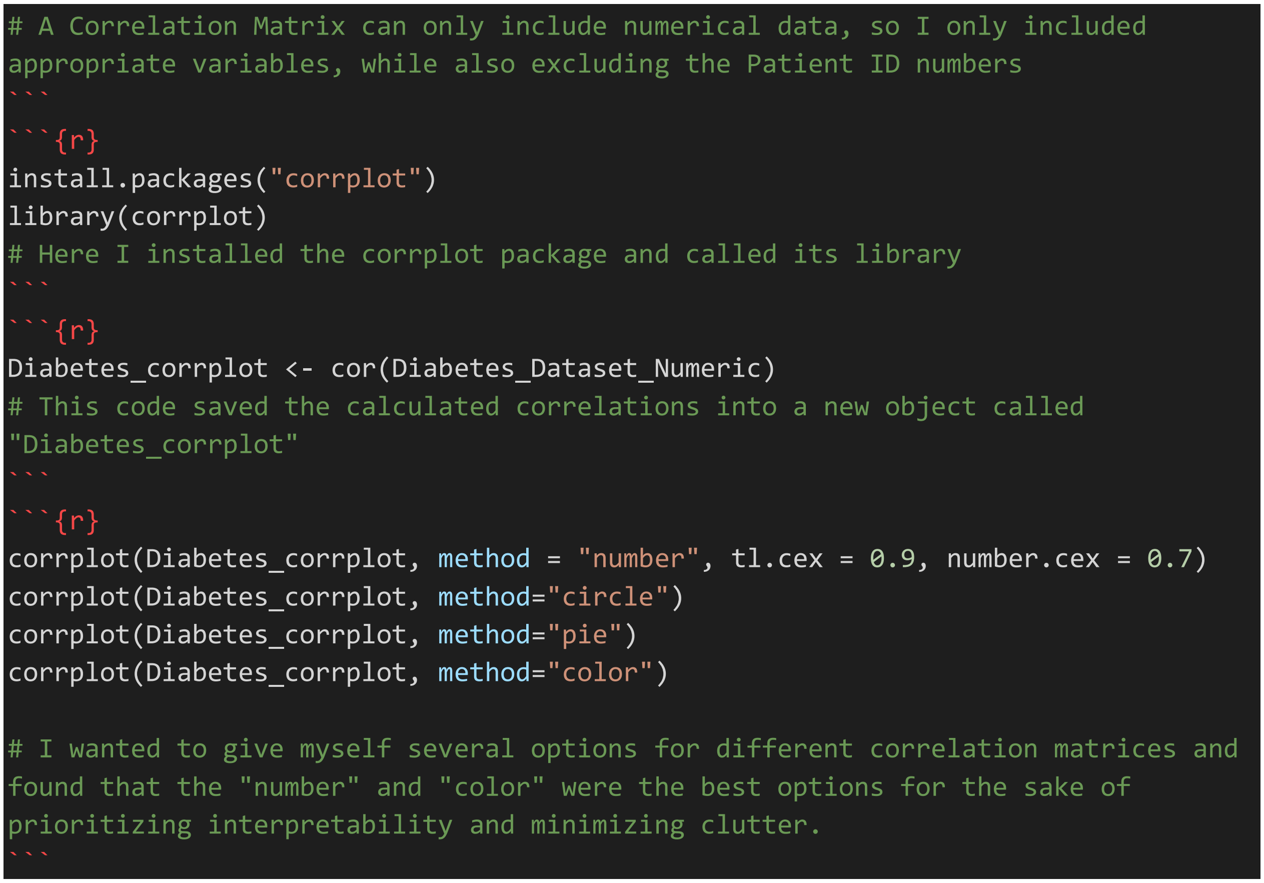
The below correlation matrix, created with a little programming in RStudio, after installing the corrplot package and loading the library, displays the values of each variable’s correlation in detail, rounded to the nearest hundredth. A correlation of 1 represents a perfect correlation, which in this case is only achieved when the variable is compared to itself, as shown in a diagonal downward, left-to-right formation across the grid. The signage of the correlation is also relevant, in that any negative value tells us that an increase in one variable’s value would result in a decrease in the other variable’s values. On the contrary, positive values represent a relationship where values of each variable should move in the same direction.

The code snippet below shows how I created the correlation matrices in RStudio:

Figure 9: R Code Used to Create Correlation Matrices

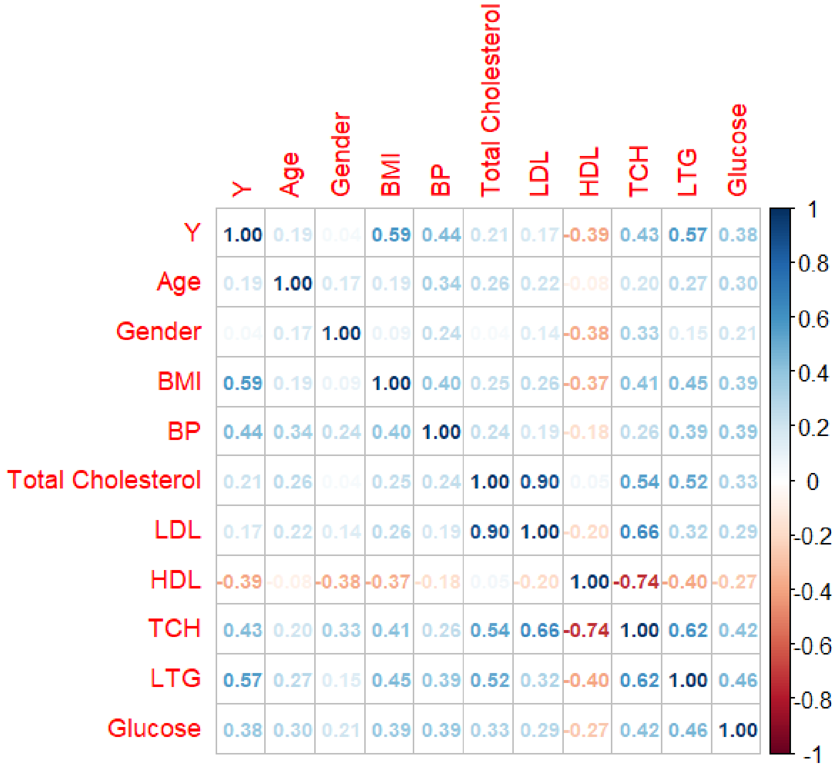


~~~
```{r}
install.packages("tidyverse")
library(dplyr)
install.packages("readxl")
library(readxl)
Diabetes_Dataset <- read_excel("Diabetes.xlsx")
# Here I am reading-in my original excel file as "Diabetes_Dataset"
```


```{r}
Diabetes_Dataset_Numeric <- Diabetes_Dataset %>%
  select_if(is.numeric) %>%
  select(-"Patient ID")
~~~



~~~
# A Correlation Matrix can only include numerical data, so I only included
appropriate variables, while also excluding the Patient ID numbers
```
```{r}
install.packages("corrplot")
library(corrplot)
# Here I installed the corrplot package and called its library
```
```{r}
Diabetes_corrplot <- cor(Diabetes_Dataset_Numeric)
# This code saved the calculated correlations into a new object called
"Diabetes_corrplot"
```
```{r}
corrplot(Diabetes_corrplot, method = "number", tl.cex = 0.9, number.cex = 0.7)
corrplot(Diabetes_corrplot, method="circle")
corrplot(Diabetes_corrplot, method="pie")
corrplot(Diabetes_corrplot, method="color")

# I wanted to give myself several options for different correlation matrices and
found that the "number" and "color" were the best options for the sake of
prioritizing interpretability and minimizing clutter.
```
~~~

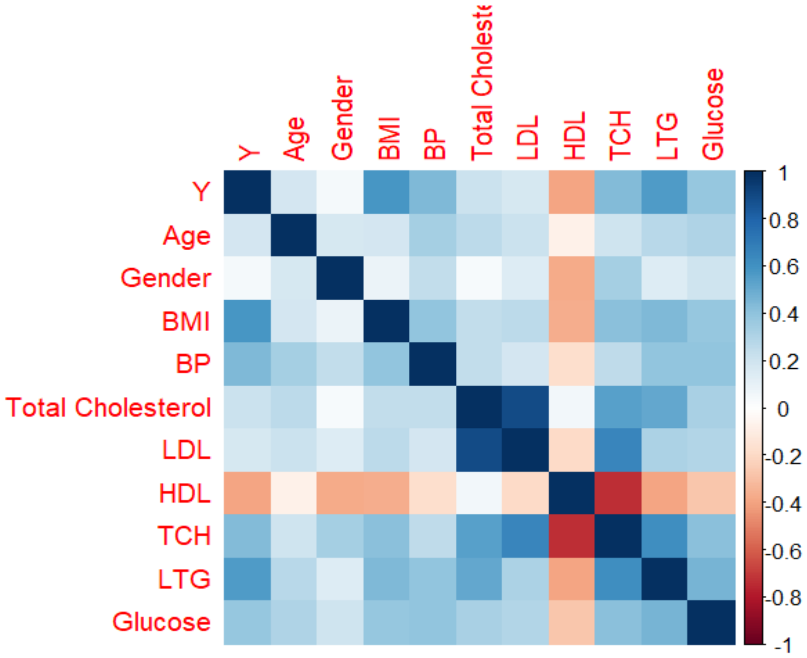
Figure 10: Correlation Matrix for Plotting Relationships between Predictor and Response Variables - Created in RStudio



	Y	Age	Gender	BMI	BP	Total Cholesterol	LDL	HDL	TCH	LTG	Glucose
Y	1.00	0.19	0.04	0.59	0.44	0.21	0.17	-0.39	0.43	0.57	0.38
Age	0.19	1.00	0.17	0.19	0.34	0.26	0.22	-0.08	0.20	0.27	0.30
Gender	0.04	0.17	1.00	0.09	0.24	0.04	0.14	-0.38	0.33	0.15	0.21
BMI	0.59	0.19	0.09	1.00	0.40	0.25	0.26	-0.37	0.41	0.45	0.39
BP	0.44	0.34	0.24	0.40	1.00	0.24	0.19	-0.18	0.26	0.39	0.39
Total Cholesterol	0.21	0.26	0.04	0.25	0.24	1.00	0.90	0.05	0.54	0.52	0.33
LDL	0.17	0.22	0.14	0.26	0.19	0.90	1.00	-0.20	0.66	0.32	0.29
HDL	-0.39	-0.08	-0.38	-0.37	-0.18	0.05	-0.20	1.00	-0.74	-0.40	-0.27
TCH	0.43	0.20	0.33	0.41	0.26	0.54	0.66	-0.74	1.00	0.62	0.42
LTG	0.57	0.27	0.15	0.45	0.39	0.52	0.32	-0.40	0.62	1.00	0.46
Glucose	0.38	0.30	0.21	0.39	0.39	0.33	0.29	-0.27	0.42	0.46	1.00

One key takeaway is made much more evident in the following heatmap, where tiles are fully shaded-in with their respective colors. In this second version, we easily see that HDL has a negative correlation with all but one of the other variables, being the "Total Cholesterol" variables, whose value partly comprises HDL:

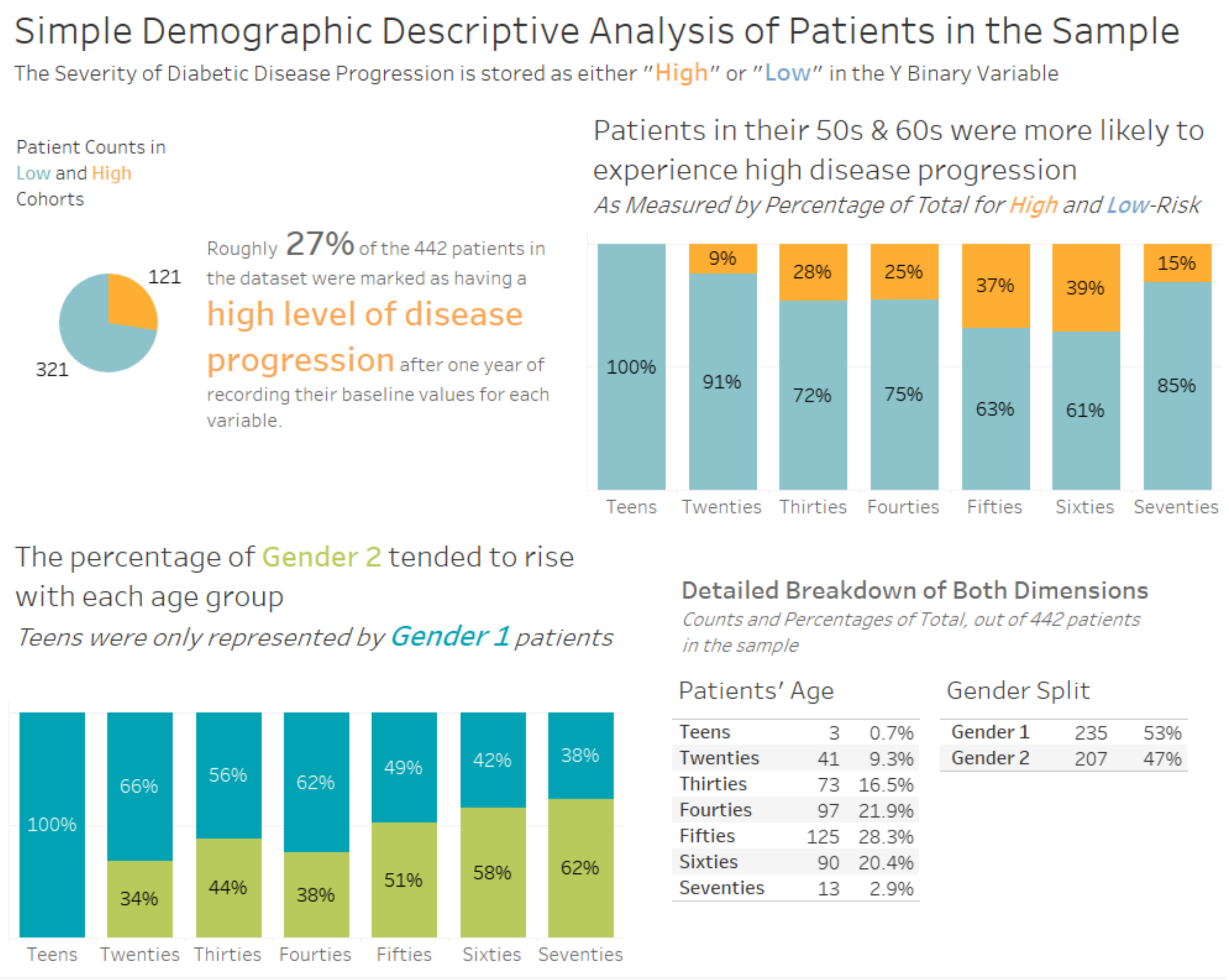
Figure 11: Heatmap of the Variable Correlations - Created in RStudio



With fully shaded tiles in the heatmap above, generalized data patterns easily arise. Unlike the first correlation matrix, however, specific relationship strengths are harder to distinguish where their shades are closer to each other. Furthermore, information about any potential visual impairments such as color blindness could result in poorer interpretability of this type of visualization, which would otherwise be fully interpretable with the first correlation matrix that includes numerical values in each tile.

As an additional layer of analysis in this project, I wanted to compare the analytic capability of JMP with Tableau, particularly in terms of each program's output and overall design. One key difference between the two lies in the cleaner appearance of Tableau. This program stands out by its simpler design by default, which is better when presenting data to an audience with a less technical background, where messages should be highlighted through the use of colors and arrows to direct their attention to the most important aspects of the visualizations. JMP, on the other hand, does not offer the option to group visualizations into a single dashboard and does not offer the option of in-depth customization. This is apparent in the following dashboards and visualizations that I produced in Tableau to explore the software's ability to display the relationship between each predictor variable in the model with the response variable of High/Low disease progression one year after baseline values were recorded.

Figure 12: Comprehensive Tableau Dashboard



Simple Demographic Descriptive Analysis of Patients in the Sample

The Severity of Diabetic Disease Progression is stored as either "High" or "Low" in the Y Binary Variable

Patient Counts in Low and High Cohorts

Roughly 27% of the 442 patients in the dataset were marked as having a high level of disease progression after one year of recording their baseline values for each variable.

Patients in their 50s & 60s were more likely to experience high disease progression

As Measured by Percentage of Total for High and Low-Risk

The percentage of Gender 2 tended to rise with each age group

Teens were only represented by Gender 1 patients

Detailed Breakdown of Both Dimensions

Counts and Percentages of Total, out of 442 patients in the sample

Patients' Age

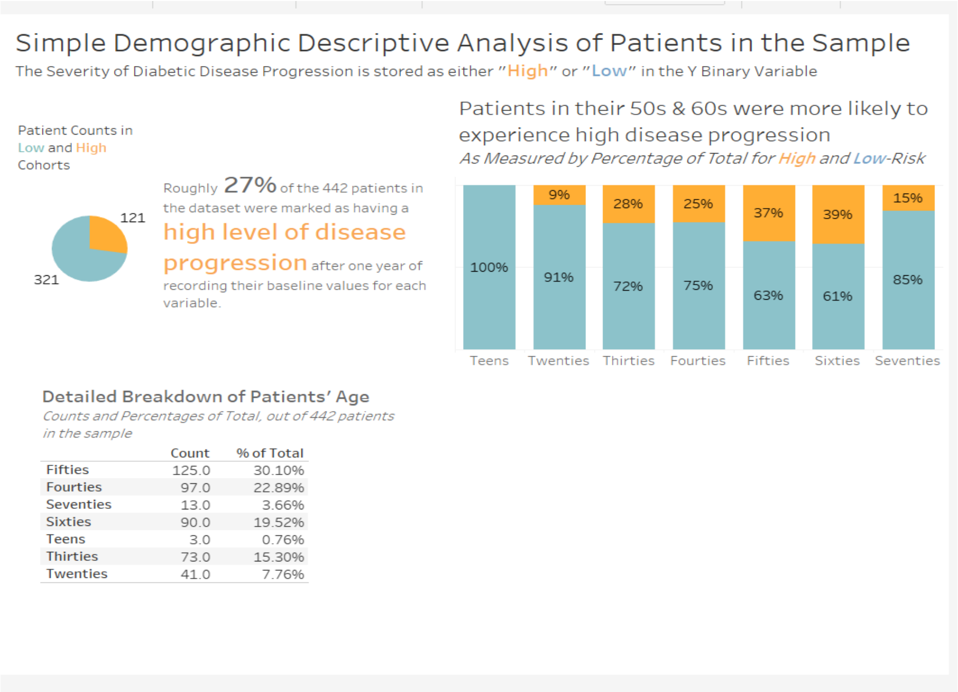
Teens	3	0.7%
Twenties	41	9.3%
Thirties	73	16.5%
Fourties	97	21.9%
Fifties	125	28.3%
Sixties	90	20.4%
Seventies	13	2.9%

Gender Split

Gender 1	235	53%
Gender 2	207	47%

The above dashboard is loaded with information which may come off as too messy and actually a hindrance to its own interpretability. So, I split the dashboard into two smaller ones based on the patients' age and the two risk groups, either "High" or "Low".

Figure 13: Tableau Dashboard Focused on Age



Simple Demographic Descriptive Analysis of Patients in the Sample

The Severity of Diabetic Disease Progression is stored as either "High" or "Low" in the Y Binary Variable

Patient Counts in Low and High Cohorts

Roughly 27% of the 442 patients in the dataset were marked as having a high level of disease progression after one year of recording their baseline values for each variable.

Patients in their 50s & 60s were more likely to experience high disease progression

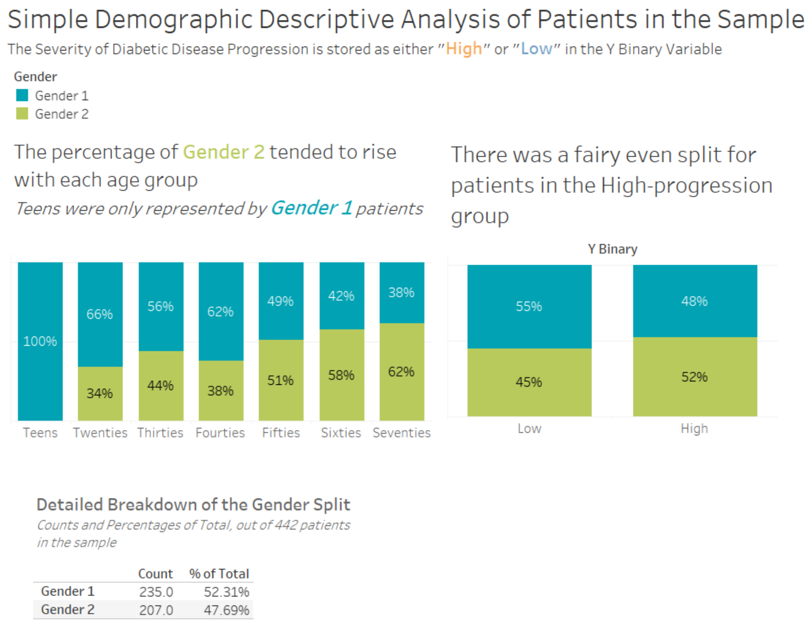
As Measured by Percentage of Total for High and Low-Risk

Detailed Breakdown of Patients' Age

Counts and Percentages of Total, out of 442 patients in the sample

	Count	% of Total
Fifties	125.0	30.10%
Fourties	97.0	22.89%
Seventies	13.0	3.66%
Sixties	90.0	19.52%
Teens	3.0	0.76%
Thirties	73.0	15.30%
Twenties	41.0	7.76%

Figure 14: Tableau Dashboard Focused on Gender



Simple Demographic Descriptive Analysis of Patients in the Sample

The Severity of Diabetic Disease Progression is stored as either "High" or "Low" in the Y Binary Variable

Gender
- Gender 1
- Gender 2

The percentage of Gender 2 tended to rise with each age group

Teens were only represented by Gender 1 patients

There was a fairy even split for patients in the High-progression group

Detailed Breakdown of the Gender Split

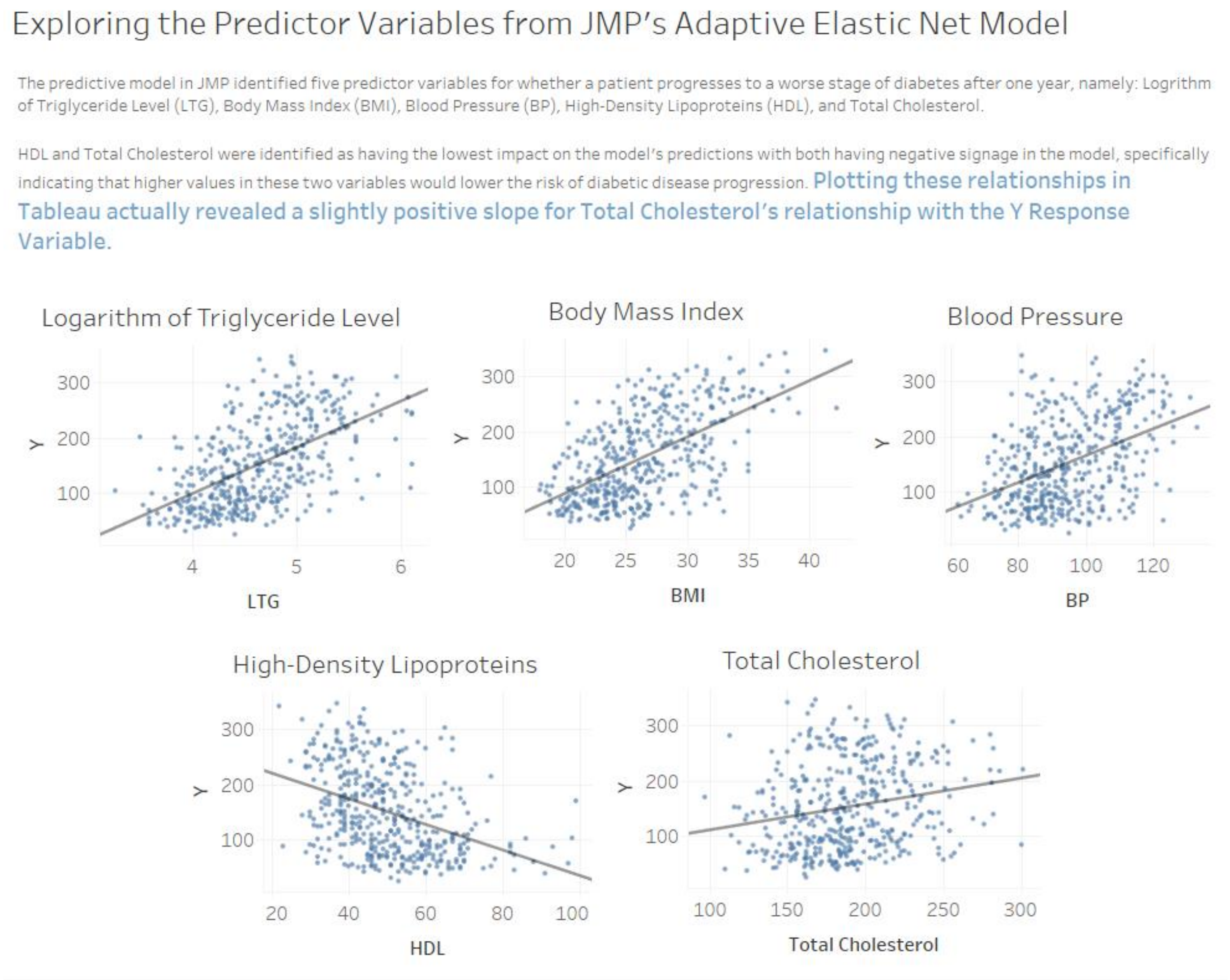
Counts and Percentages of Total, out of 442 patients in the sample

	Count	% of Total
Gender 1	235.0	52.31%
Gender 2	207.0	47.69%

I wanted to explore the statistical capabilities of Tableau, as a means to compare the output and design with that of JMP. Tableau is overall more attractive as a software with can be helpful when presenting results to an audience that may not have much experience in analyzing data, and therefore will need to rely on simple graphs with clear, easy-to-follow headings, and additional help with color and symbols to guide them into seeing the important messages that each visualization conveys. In the two simplified versions, takeaways are clearly defined in the headers, and subtitles clarify any potential for confusion. Additionally, tables are provided with exact numerical figures for patient counts and percentage-of-total values in the two dimensions.

With the series of scatter plots below, my intent is to display a simple relationship between each 5 predictor variables that were included in my final JMP model, with the response variable (the column with the numerical equivalent to the binary Y column, where any value greater than 200 constituted "High" diabetic disease progression):

Figure 15: Dashboard for Scatterplots Produced in Tableau



Exploring the Predictor Variables from JMP's Adaptive Elastic Net Model

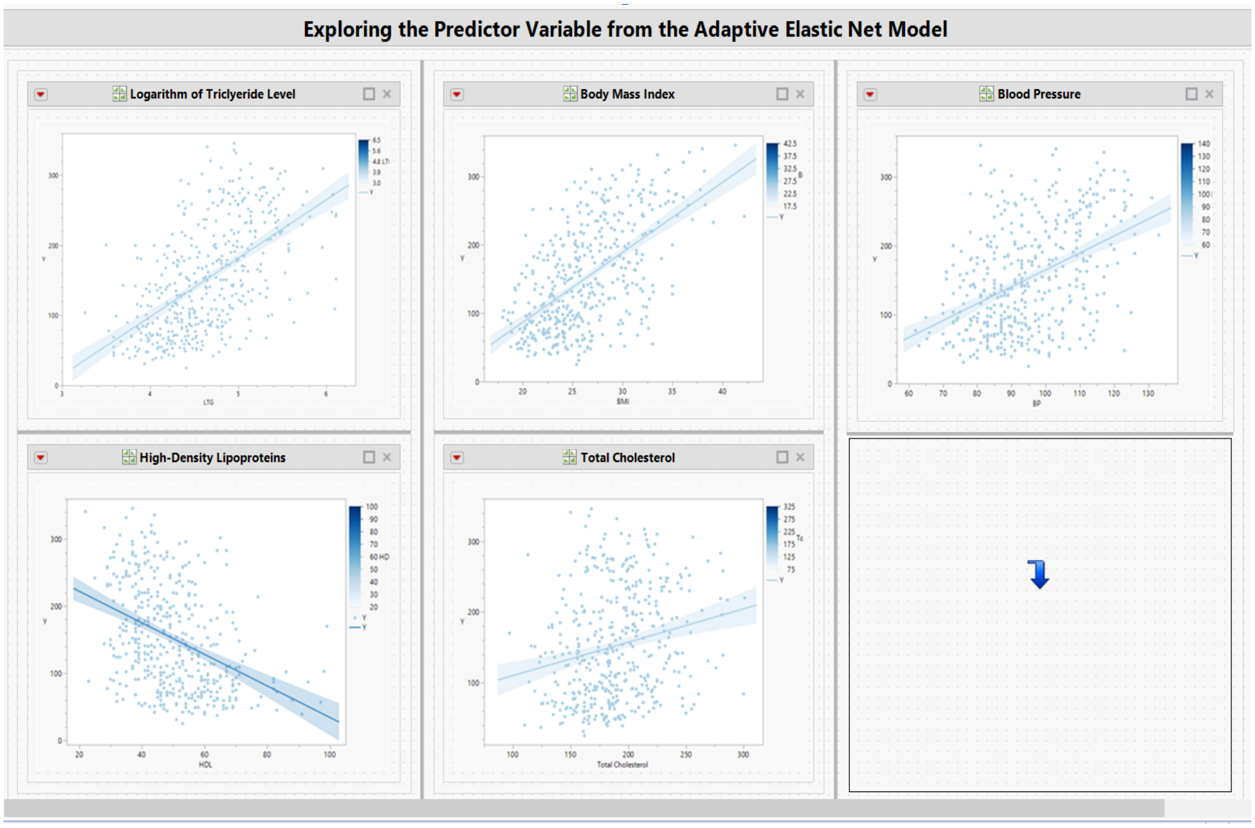
The predictive model in JMP identified five predictor variables for whether a patient progresses to a worse stage of diabetes after one year, namely: Logrithm of Triglyceride Level (LTG), Body Mass Index (BMI), Blood Pressure (BP), High-Density Lipoproteins (HDL), and Total Cholesterol.

HDL and Total Cholesterol were identified as having the lowest impact on the model's predictions with both having negative signage in the model, specifically indicating that higher values in these two variables would lower the risk of diabetic disease progression. **Plotting these relationships in Tableau actually revealed a slightly positive slope for Total Cholesterol's relationship with the Y Response Variable.**

While Tableau produces graphs that are very visually appealing, yet still emphasizing their core, analytical purpose, scatter plots are a type of visualization that both JMP and RStudio are capable of reproducing as well. As I will show in the following figures, Tableau outperforms the other programs in how more data can be fit into a compact, yet attractive dashboard, while the other programs are much more focused on the practicality of the plots.

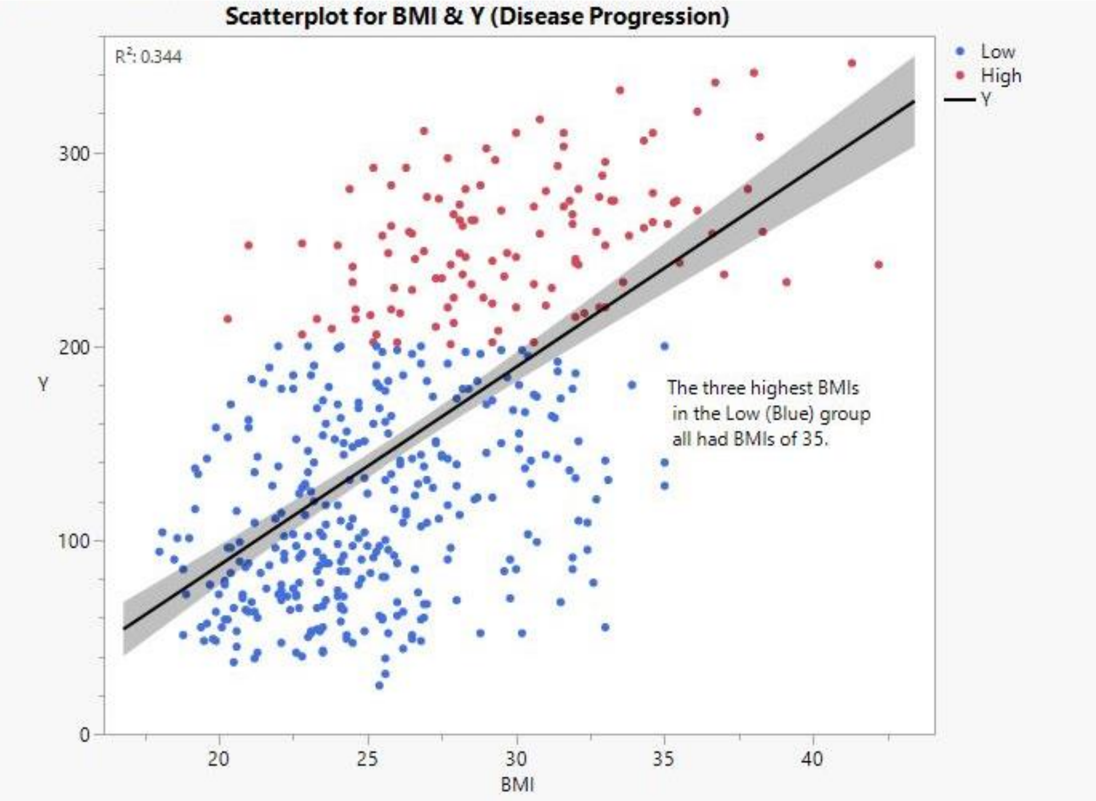
To illustrate this point, I used JMP's dashboard building tool to try to replicate the one that I created in Tableau, in figure 15:

Figure 16: Dashboard for Scatterplots Produced in JMP



The Dashboard building tool within JMP is much less oriented towards creating visually appealing insights into the data, as it is for displaying a clear focus on the data. JMP's dashboard builder has a small series of templates to choose from with semi-flexible and no specific options for 5 visualizations of the same dimensions, unless I were to dabble with the "Blank Dashboard" option. Also, visualizations within the dashboard cannot be treated as floating objects, like in Tableau, which eliminates the ability to use white space to my advantage and replicate my previous design where the bottom two visualization can be centered in their row. Additionally, labels cannot be altered in terms of color or font style, while also not giving the option to hide the rest of the elements in the header bar (the red drop-down arrows, the green graph-builder icon, and the maximize icon), especially since these icon cannot be toggled with from within the Dashboard. Overall, JMP is an extremely powerful application for predictive analytics, but it cannot be seen as a substitute for Tableau when it comes to incorporating a more creative approach to displaying data, which is particularly important when the audience is not familiar with the data and is much more reliant on visual cues and design choices to quickly digest what they are seeing. The boxiness of JMP's dashboards and the non-functional icons included in the scatterplots' headers create unnecessary noise within the dashboards that inhibits its interpretability in this case.

Figure 17: Scatterplot for BMI and Disease Progression - Created in JMP

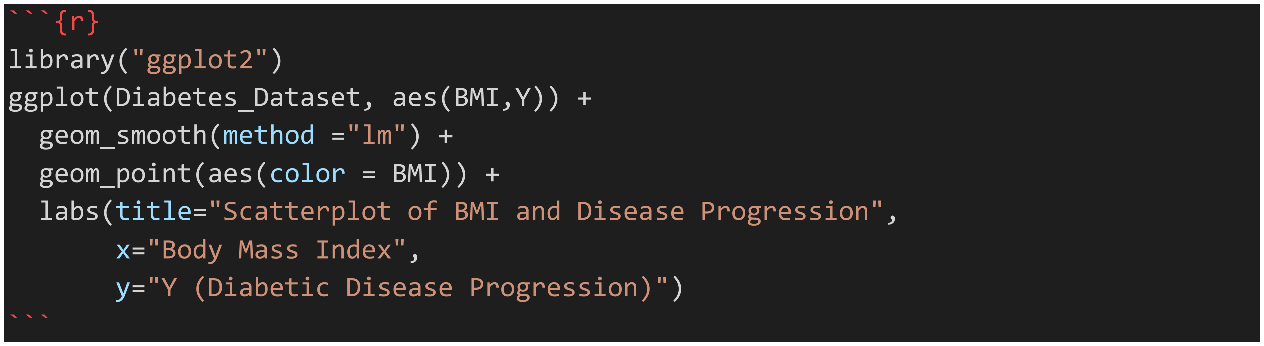


With less focus on matching the overall color-theme in a dashboard comprising many different scatterplots, creating one within JMP with a focus on only BMI led me to be more open to play with bolder colors for the data points, where Red and Blue represents the High and Low disease progression patients from the binary response variable, respectively. We can see by the overall shape of the data points that disease progression levels tended to increase with BMIs, further supported by the positive slope of the black Line of Fit. While there is a large degree of dispersion for the highest BMIs, there were no BMIs within the "Low" group that exceeded a value of 35, while there were many patients in the "High" group that had baseline BMI values that were greater than 35.

In terms of interpretability, Tableau is a better choice when wanting to include a series of scatterplots onto a single dashboard without trading off the quality of each individual scatterplot. Below is an example of how JMP Dashboard would appear by creating a static dashboard to simply view all 5 scatterplots for the predictor variables that were selected by JMP's adaptive elastic net model:

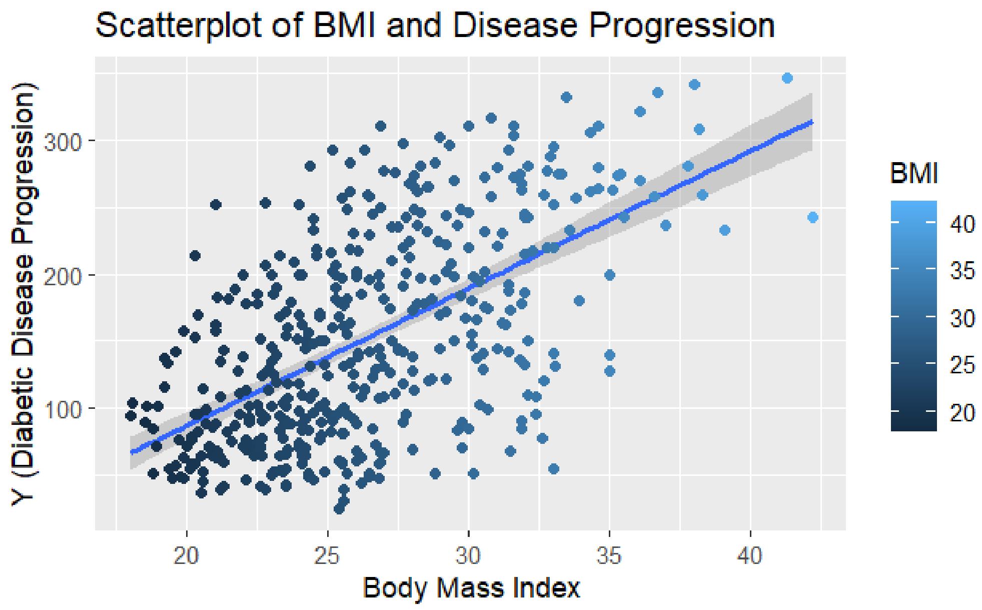
Additionally, the following figure contains the necessary code in R to create the equivalent scatterplot for BMIs and values in the Y response variable within RStudio. To create this scatterplot, I loaded the "ggplot2" library, which was installed within the tidyverse package shown previously in figure 9.

Figure 18: R Code Used to Create Scatterplot



~~~
```{r}
library("ggplot2")
ggplot(Diabetes_Dataset, aes(BMI,Y)) +
  geom_smooth(method ="lm") +
  geom_point(aes(color = BMI)) +
  labs(title="Scatterplot of BMI and Disease Progression",
       x="Body Mass Index",
       y="Y (Diabetic Disease Progression)")
```
~~~

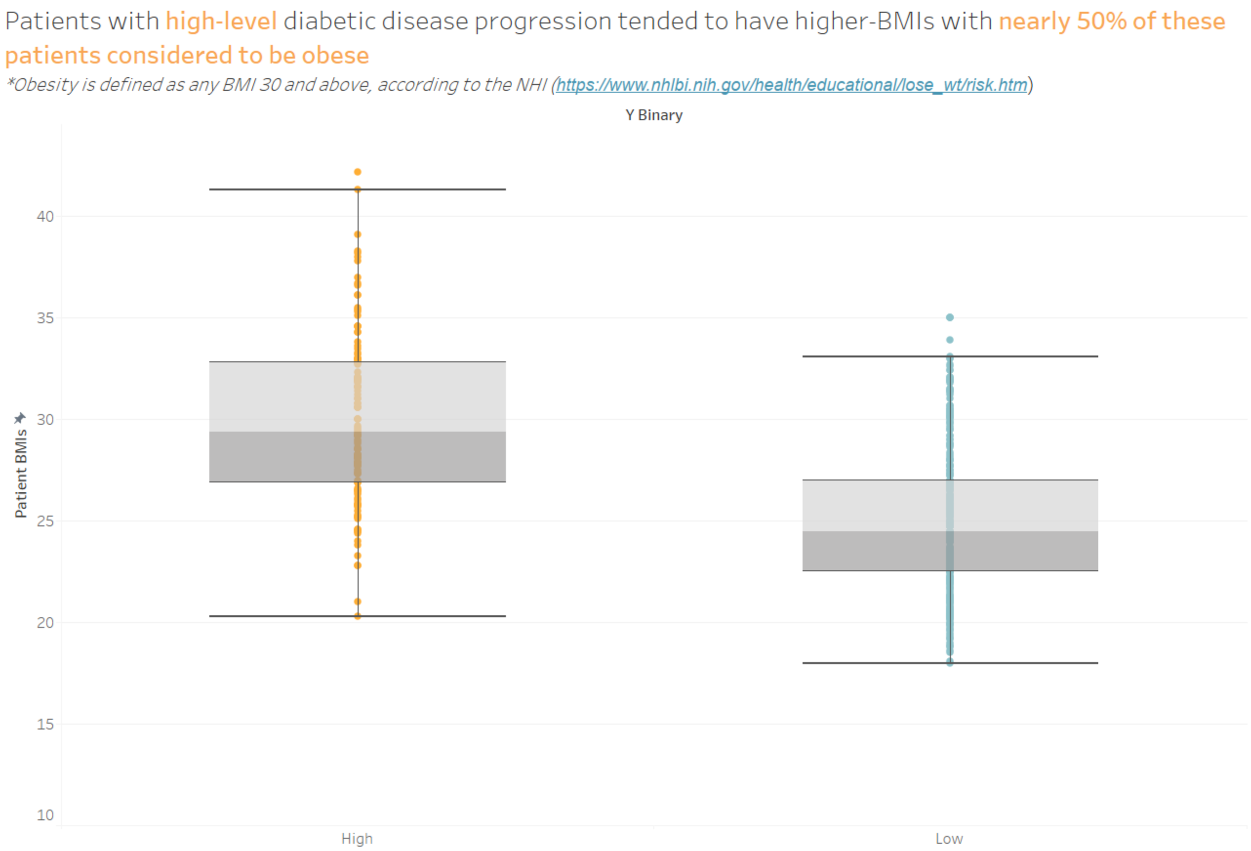
Figure 19: Scatterplot of BMI and Disease Progression Created in RStudio



From this scatterplot, we can see that the output is rather boxy and much more data-centric, rather than allowing a higher degree of design choices that would stray beyond color, labels, and font size. Dashboards are capable of being built in RStudio with the installation of the "flexdashboard" package, but I will omit any additional version of the same dashboard that I previously created in Tableau and JMP since there is no analytical benefit in doing so. For the sake of its interpretability for the audience, high scalability, and ease-of-use for the user, I would have to endorse Tableau as the better of the two programs when producing dashboards.

This next infographic includes two histograms to visualize the distribution of patient BMIs in each disease progression group, given that this variable was determined to be a very important predictor variable in the model. The story I wanted to highlight from my analysis using Tableau is that nearly 50% of patients in the response variable group that experienced high disease progression after 1 year were initially obese, as determined by having a BMI of at least 30. Much like the scatterplot for BMIs and the Y Response Variable values from JMP (see figure 16), I incorporated an additional layer of information within the visualization by encoding the two groups corresponding to High and Low levels of disease progression with color within the title to eliminate the need for a legend to take up more space, where orange is representative of the "High" (disease progression) group:

Figure 20: Exploring BMI in the High Disease Progression Group with Boxplots

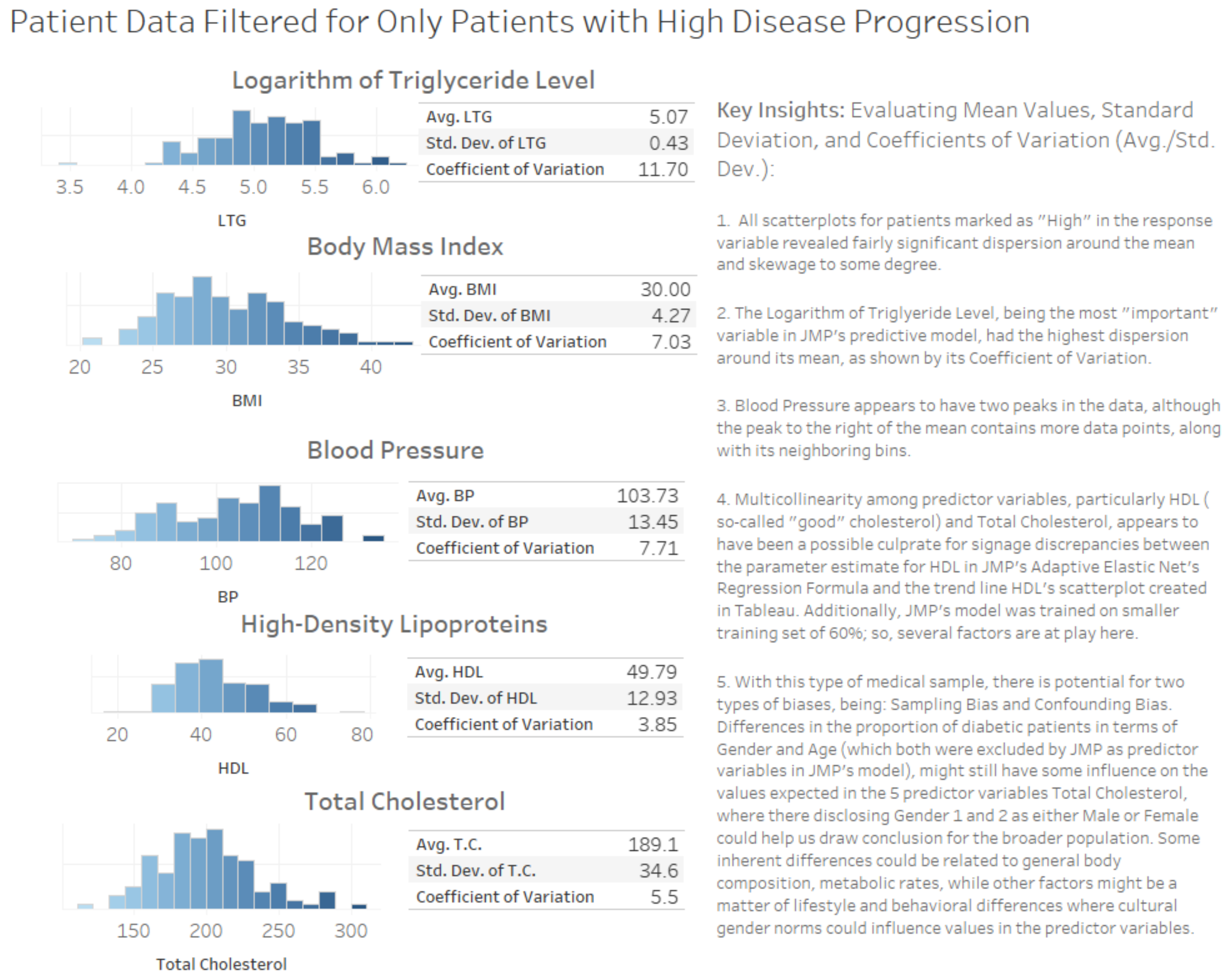


Patients with high-level diabetic disease progression tended to have higher-BMIs with nearly 50% of these patients considered to be obese

*Obesity is defined as any BMI 30 and above, according to the NHI (https://www.nhlbi.nih.gov/health/educational/lose_wt/risk.htm)

Other key takeaways were included in the following dashboard that I created in Tableau after filtering the data for only patients in the High disease progression group, unlike the previously shared histograms produced in JMP. Here, some of the visual aspects to focus on are relative to the general functionality of histograms in how they display the distribution of the data points and any skewness or outliers present, as well as the number of peaks in the data. I also included a calculated field for the Coefficient of Variation in each table to capture the level of dispersion around the mean values, as a ratio of the standard deviation over the mean.

*Figure 21: Further Analyzing Predictor Variables with Histograms in Tableau**



Patient Data Filtered for Only Patients with High Disease Progression

Logarithm of Triglyceride Level

Avg. LTG	5.07
Std. Dev. of LTG	0.43
Coefficient of Variation	11.70

Body Mass Index

Avg. BMI	30.00
Std. Dev. of BMI	4.27
Coefficient of Variation	7.03

Blood Pressure

Avg. BP	103.73
Std. Dev. of BP	13.45
Coefficient of Variation	7.71

High-Density Lipoproteins

Avg. HDL	49.79
Std. Dev. of HDL	12.93
Coefficient of Variation	3.85

Total Cholesterol

Avg. T.C.	189.1
Std. Dev. of T.C.	34.6
Coefficient of Variation	5.5

Key Insights: Evaluating Mean Values, Standard Deviation, and Coefficients of Variation (Avg./Std. Dev.):

1. All scatterplots for patients marked as "High" in the response variable revealed fairly significant dispersion around the mean and skewage to some degree.

2. The Logarithm of Triglyeride Level, being the most "important" variable in JMP's predictive model, had the highest dispersion around its mean, as shown by its Coefficient of Variation.

3. Blood Pressure appears to have two peaks in the data, although the peak to the right of the mean contains more data points, along with its neighboring bins.

4. Multicollinearity among predictor variables, particularly HDL (so-called "good" cholesterol) and Total Cholesterol, appears to have been a possible culprate for signage discrepancies between the parameter estimate for HDL in JMP's Adaptive Elastic Net's Regression Formula and the trend line HDL's scatterplot created in Tableau. Additionally, JMP's model was trained on smaller training set of 60%; so, several factors are at play here.

5. With this type of medical sample, there is potential for two types of biases, being: Sampling Bias and Confounding Bias. Differences in the proportion of diabetic patients in terms of Gender and Age (which both were excluded by JMP as predictor variables in JMP's model), might still have some influence on the values expected in the 5 predictor variables Total Cholesterol, where there disclosing Gender 1 and 2 as either Male or Female could help us draw conclusion for the broader population. Some inherent differences could be related to general body composition, metabolic rates, while other factors might be a matter of lifestyle and behavioral differences where cultural gender norms could influence values in the predictor variables.

**Here Key Insight 4 should state that there are discrepancies in signage between the parameter estimate for Total Cholesterol from JMP's model's prediction formula (figure 7 in this report) and direction of the corresponding slope for Total Cholesterol produced (see figure 15).*

In conclusion, Tableau provides analysts with the opportunity for a higher level of personalization when producing visualizations. The primary focus when working within this program is on creating descriptive reports, where key dashboards and stories (not shown in this report) allow users to share several findings on a single page. When producing compelling and effective dashboards, analysts can leverage the power of Pre-attentive Attributes and the Gestalt Principles of Visual Perception (through color, font, italicized text, etc.) to ensure that their visualizations are easily interpretable and compelling for the target audience. As for JMP, this program provides a much simpler user interface that makes the process of building predictive models more intuitive. Alternatively, predictive models can be constructed in RStudio using the R programming language with the right packages and with some limitations in Tableau with the use of external tools, particularly as it pertains to complex models like the Adaptive Elastic Net used in this report, which is not possible to build natively within Tableau. That being said, results can be saved into a compatible file and then uploaded to Tableau to build visualization after performing these building these models in other applications with stronger predictive capabilities, such as JMP or RStudio. So, while my conclusions drawn from the Adaptive Elastic Model's prediction profilers and selected variables of importance could be helpful in determining how each patient's baseline values may have impacted their health, no medical implications should be extrapolated for the entire population of diabetics, nor should variables dropped when constructing the predictive model be disregarded when discussing one's risk. This report purely serves as a comparative analysis of the applications of JMP, Tableau, and R, in how they can be used to analyze the relationship between variables within a dataset for both descriptive and predictive analytics.

References

“Assessing Your Weight.” *Centers for Disease Control and Prevention*, Centers for Disease Control and Prevention, 3 June 2022, www.cdc.gov/healthyweight/assessing/index.html#:~:text=If%20your%20BMI%20is%20less,falls%20within%20the%20obese%20range.

Efron, Bradley, et al. “Least angle regression.” *The Annals of Statistics*, vol. 32, no. 2, 2004, https://doi.org/10.1214/009053604000000067, https://projecteuclid.org/journals/annals-of-statistics/volume-32/issue-2/Least-angle-regression/10.1214/009053604000000067.full.

“High Blood Triglycerides.” *National Heart Lung and Blood Institute*, U.S. Department of Health and Human Services, www.nhlbi.nih.gov/health/high-blood-triglycerides. Accessed 14 Sept. 2023.

Nakrani, Mihir N., et al. “Physiology, Glucose Metabolism - Statpearls - NCBI Bookshelf.” *National Library of Medicine*, National Institutes of Health, 17 July 2023, www.ncbi.nlm.nih.gov/books/NBK560599/.

“Understanding Blood Pressure Readings.” *Www.Heart.Org*, 30 May 2023, www.heart.org/en/health-topics/high-blood-pressure/understanding-blood-pressure-readings.

Table of Figures:

(Left click on the titles to navigate to each figure)