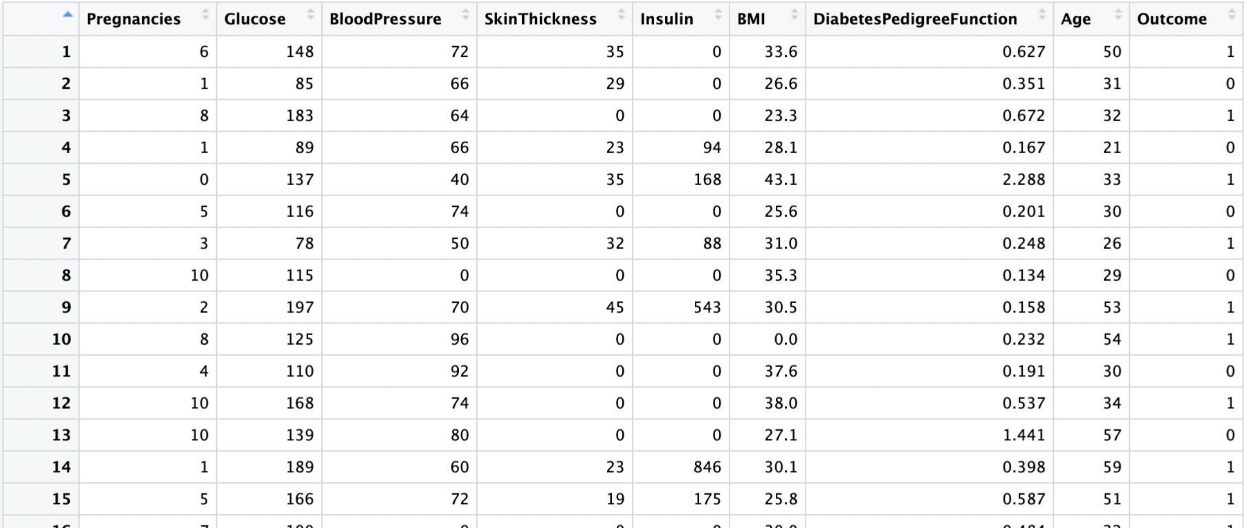
Using R to explore the diabetes dataset

Introduction

This dataset is originally from the National Institute of Diabetes and Digestive and Kidney Diseases. The objective of the dataset is to diagnostically predict whether or not a patient has diabetes, based on certain diagnostic measurements included in the dataset. Several constraints were placed on the selection of these instances from a larger database. In particular, all patients here are females at least 21 years old of Pima Indian heritage. It consists of several medical predictor variables and one target variable,outcome. Predictor variables include the number of pregnancies the patient has had, their BMI, insulin level, age, and so on.

In this report, I will accomplish 4 analytical objectives: visualization, descriptive statistics, linear modeling, and hypothesis testing based on this diabetes dataset.

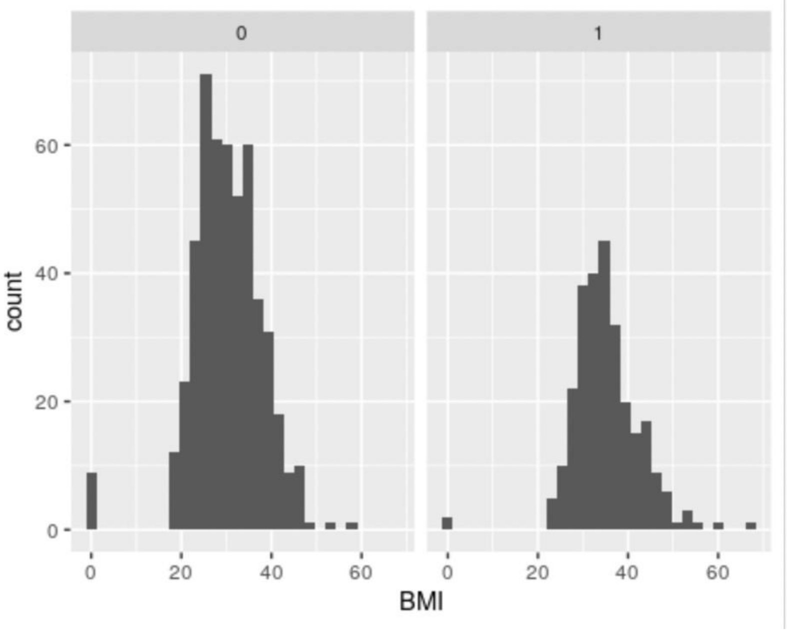


	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age	Outcome
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0
7	3	78	50	32	88	31.0	0.248	26	1
8	10	115	0	0	0	35.3	0.134	29	0
9	2	197	70	45	543	30.5	0.158	53	1
10	8	125	96	0	0	0.0	0.232	54	1
11	4	110	92	0	0	37.6	0.191	30	0
12	10	168	74	0	0	38.0	0.537	34	1
13	10	139	80	0	0	27.1	1.441	57	0
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1

Visualizations

Histogram of BMI

```
ggplot(diabetes, aes(x=BMI))+
  geom_histogram()+facet_wrap(~Outcome)
```

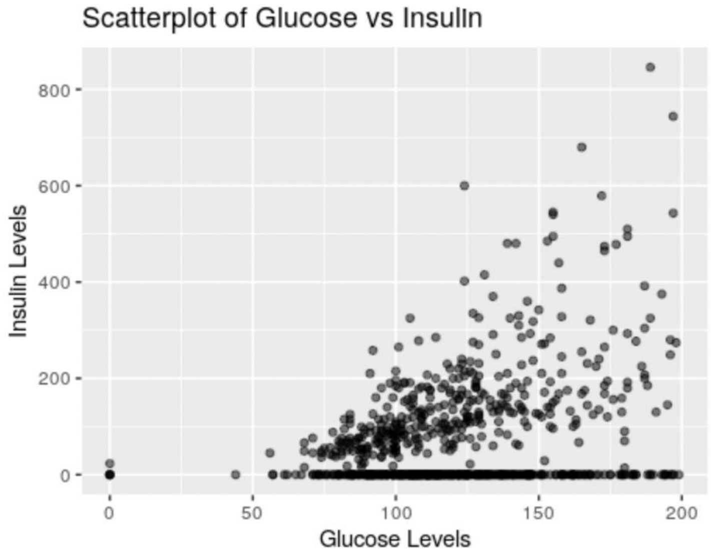


Interpret:

For both groups, the BMI distribution appears relatively normally distributed, while the diabetes group centered around a mean BMI in the high 30s
Diabetic patients appear to have higher BMIs on average compared to non-diabetic patients
The BMI distribution for diabetic patients also appears to be more skewed to the right with heavier tails, indicating higher variability of BMIs

Scatterplot of glucose vs insulin

```
ggplot(data = diabetes, aes(x = Glucose, y = Insulin)) +
  geom_point(alpha = 0.5) +
  labs(
    title = "Scatterplot of Glucose vs Insulin",
    x = "Glucose Levels",
    y = "Insulin Levels"
  )
```



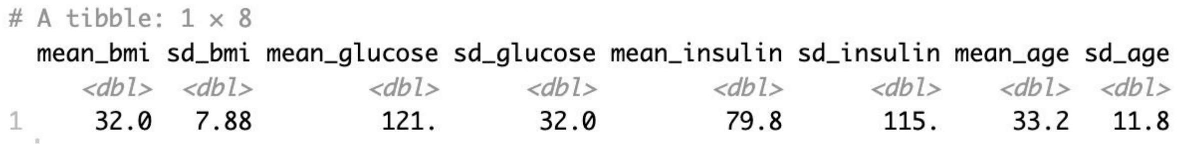
Interpret:

The scatterplot shows the relationship between blood glucose levels and blood insulin levelsfor the patients in the dataset. There is a moderately strong positive correlation between glucose and insulin levels. As glucose levels increase, insulin levels also tend to increase. This indicates that patients with higher blood glucose levels also tend to have higher insulin levels.

Descriptive Statistics

Summary tables of mean, SD for numeric variables like BMI, glucose, insulin, age

```
diabetes %>%
 summarize(
  mean_bmi = mean(BMI, na.rm = TRUE),
  sd_bmi = sd(BMI, na.rm = TRUE),
  mean_glucose = mean(Glucose, na.rm = TRUE),
  sd_glucose = sd(Glucose, na.rm = TRUE),
  mean_insulin = mean(Insulin, na.rm = TRUE),
  sd_insulin = sd(Insulin , na.rm = TRUE),
  mean_age = mean(Age, na.rm = TRUE),
  sd_age = sd(Age, na.rm = TRUE)
 )
```



```
# A tibble: 1 × 8
  mean_bmi sd_bmi mean_glucose sd_glucose mean_insulin sd_insulin mean_age sd_age
     <dbl>  <dbl>        <dbl>      <dbl>        <dbl>      <dbl>    <dbl>  <dbl>
1     32.0   7.88         121.       32.0         79.8       115.     33.2   11.8
```

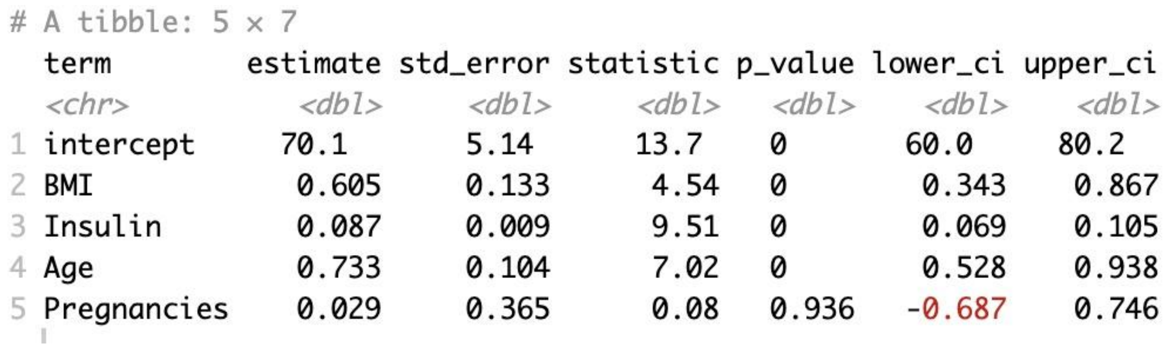
Interpret:

The average BMI in the dataset is 32, which falls into the obese BMI range between 30-35. Mean blood glucose levels are very high at 121 mg/dL on average. A normal range would be under 100 mg/dL fasting. The standard deviation is 32, so there is a substantial spread of glucose levels amongst the patients.
Average insulin level is 79.8 pmol/L. I can't assess full variation due to the zeros, but the non-zero insulin values have large dispersion as evidenced by the standard deviation of 115 pmol/L.
Mean age is 33.2 years with a standard deviation of 11.8 years. So the typical patient age is early 30s, with ages in the dataset ranging from at least 21 up beyond middle age.

Linear Model

Linear model predicting glucose levels from BMI, insulin, age, # pregnancies

```
model <- lm(Glucose ~ BMI + Insulin + Age + Pregnancies, data=diabetes)
get_regression_table(model)
```



```
# A tibble: 5 × 7
  term        estimate std_error statistic p_value lower_ci upper_ci
  <chr>          <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
1 intercept     70.1       5.14      13.7    0        60.0     80.2
2 BMI            0.605     0.133      4.54   0         0.343    0.867
3 Insulin        0.087     0.009      9.51   0         0.069    0.105
4 Age            0.733     0.104      7.02   0         0.528    0.938
5 Pregnancies    0.029     0.365      0.08   0.936    -0.687    0.746
```

Interpret:

BMI has a positive coefficient of 0.605,which means that for every 1 unit increase in BMI, glucose levels increase by an estimated 0.605 mg/dL on average, this effect is statistically significant based on the low p-value and confidence interval not crossing 0.
Insulin also has a significant positive estimated effect of 0.087. So glucose tends to increase by 0.087 mg/dL for each 1 unit rise in insulin.
Age has a significant coefficient of 0.733. This means the patient's glucose is higher on average by 0.733 mg/dL for each additional year of age, controlling for the other variables.
Pregnancies value does not have any effect - the p-value is > 0.05 and confidence interval includes 0, meaning this test cannot conclude a non-zero effect based on this model and data.
In summary, higher BMI, insulin, and age are significantly associated with increased glucose levels independently based on this multivariate regression analysis.

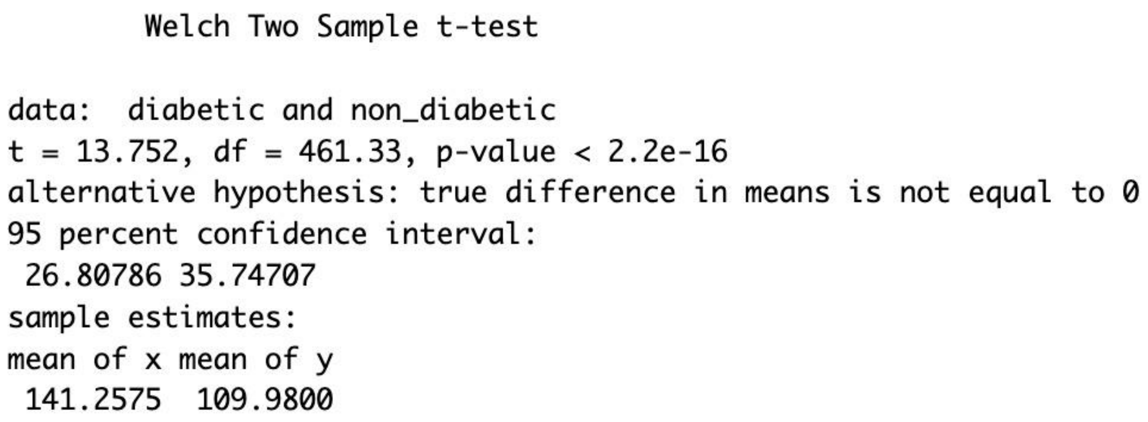
Hypothesis Testing

Hypothesis

H0: There is no difference in mean glucose levels between diabetic (Outcome=1) and non-diabetic groups
H1: Diabetic patients have higher mean glucose on average

```
diabetic <- subset(diabetes, Outcome==1)$Glucose
non_diabetic <- subset(diabetes, Outcome==0)$Glucose

t.test(diabetic, non_diabetic, var.equal = FALSE)
```



```
        Welch Two Sample t-test

data:  diabetic and non_diabetic
t = 13.752, df = 461.33, p-value < 2.2e-16
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 26.80786 35.74707
sample estimates:
mean of x mean of y
 141.2575  109.9800
```

Interpret:

The extremely small p-value < 2.2e-16 means there is incredibly strong evidence to reject the null hypothesis. There is a statistically significant difference in average glucose levels between the diabetic and non-diabetic groups.
Since diabetics have impaired glucose control, the alternative hypothesis states their levels should be higher. So based on this highly significant test result, it can infer that diabetic patients have substantially higher glucose levels on average compared to non-diabetic patients in the dataset.
The 95% confidence interval also supports this conclusion. In summary, the t-test confirms diabetic patients display markedly higher blood glucose compared to patients without diabetes, aligning with physiological understanding.

Conclusion

The analysis of the diabetes dataset reveals that diabetic patients, characterized by higher BMI, insulin levels, and age, exhibit significantly elevated blood glucose levels compared to non-diabetic patients. This is supported by strong evidence from visualizations, descriptive statistics, a linear model, and a hypothesis test, emphasizing the robust association between certain patient characteristics and diabetes-related outcomes.