

INF 2178 Final

Professor: Shion Guha

Student: Huimin Yvonne Zhan

Student ID: 1007554834

Goal

By applying the t-test and ANOVA test, this project aims to learn how Glucose, BMI, and Insulin correlate with the possibility of Diabetes II.

Introduction

According to earlier research indicated, the rising type 2 diabetes is not only crucial for individual health status, but also a chronic and 'genetic' disease with certain probability passing to our offspring and challenging 'health economy' (Tine D. Clausen; Elisabeth R. Mathiesen; Torben Hansen; Oluf Pedersen; Dorte M. Jensen; Jeannet Lauenborg; Peter Damm, 2008).

Thus, it's worthwhile for us to conduct an in-depth research to see what's the key factor(s) causing the rising number of type 2 diabetes. Based on the Diabetes.csv from the National Institute of Diabetes & Digestive & Kidney Diseases, we briefly saw there are 8 major different variances relating to the possibility of having diabetes, including BMI, Glucose, BloodPressure, Diabetes Pedigree Function, Insulin, Pregnancies and SkinThickness. All of those 8 different variances correlate with each other to certain degrees. The dataset offers a binary result of Diabetes showing whether or not a patient has diabetes with those 8 numerical variances impacts. Thus, we did an in-depth research to explore how top 3 variances correlatej. to each other, including Glucose, BMI and Insulin. Through applying ANOVA AND t-test to compare those three variance, we learnt that Glucose is the dependent variable, BMI and Insulin are independent variables in this case.

Method

Since there is no research problem indicated at the 'diabetes.csv' and outlined in this assignment, we cleaned the dataset and took an heatmap to visualize the top 3 variances before conducting the correlation research.

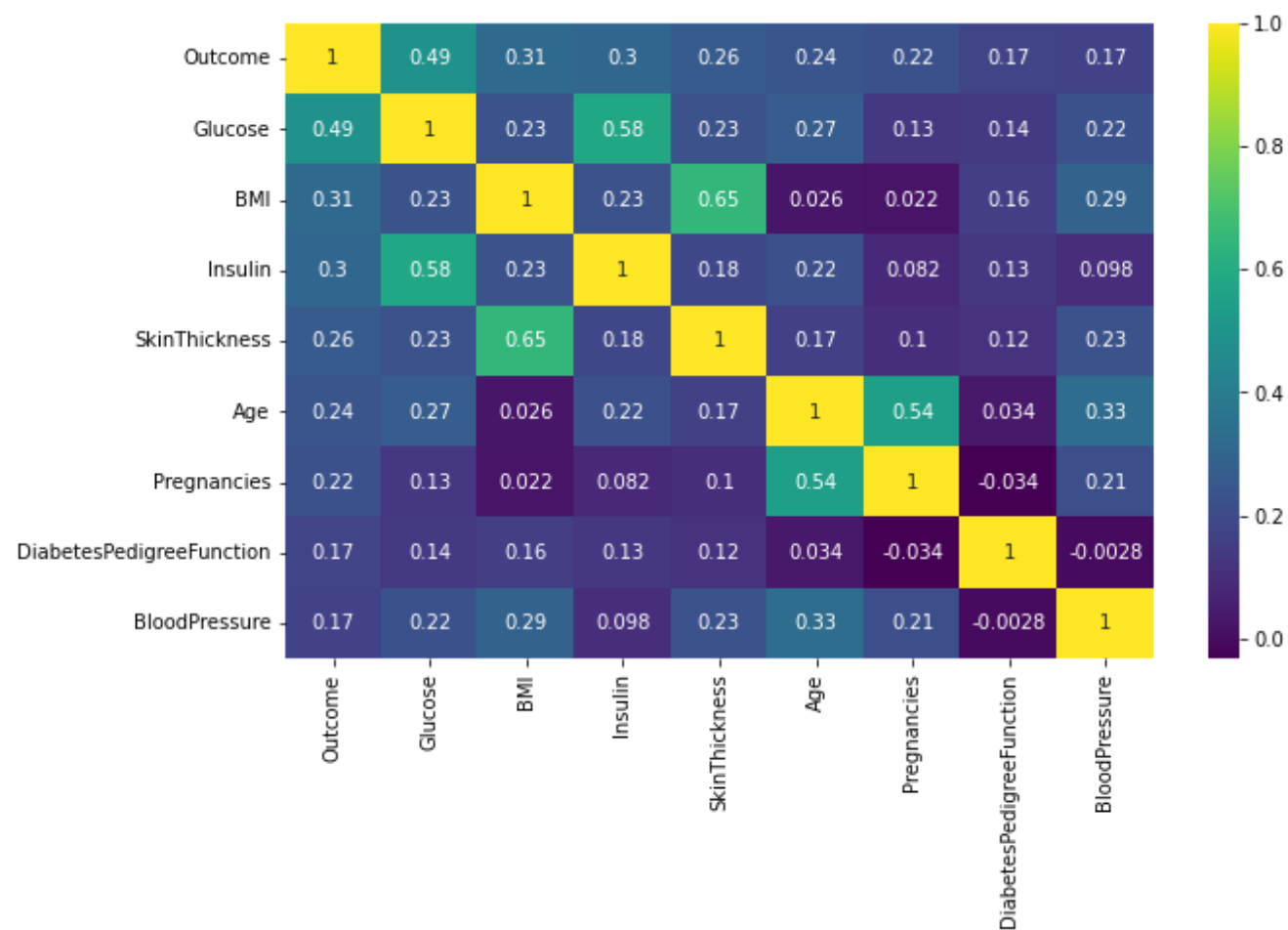
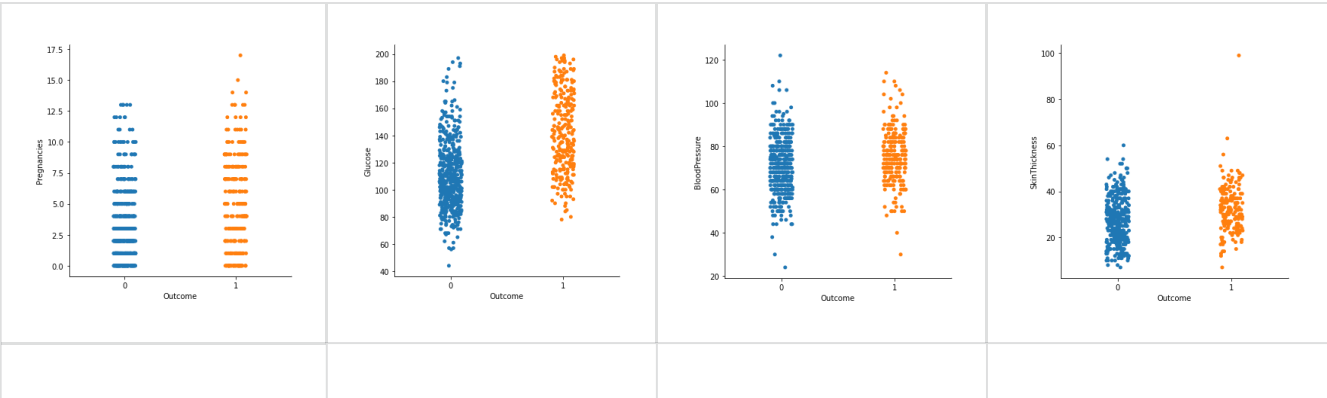
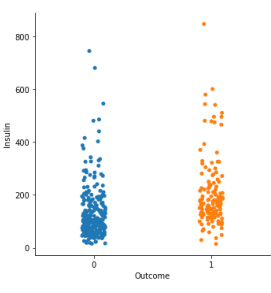
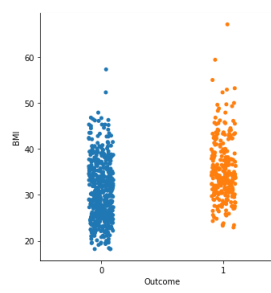
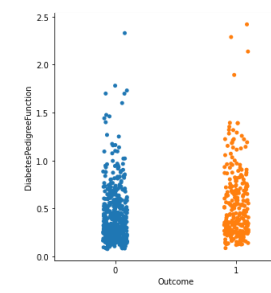
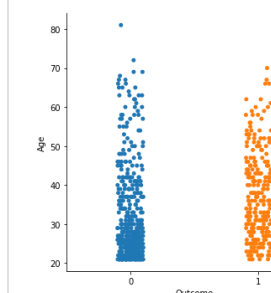


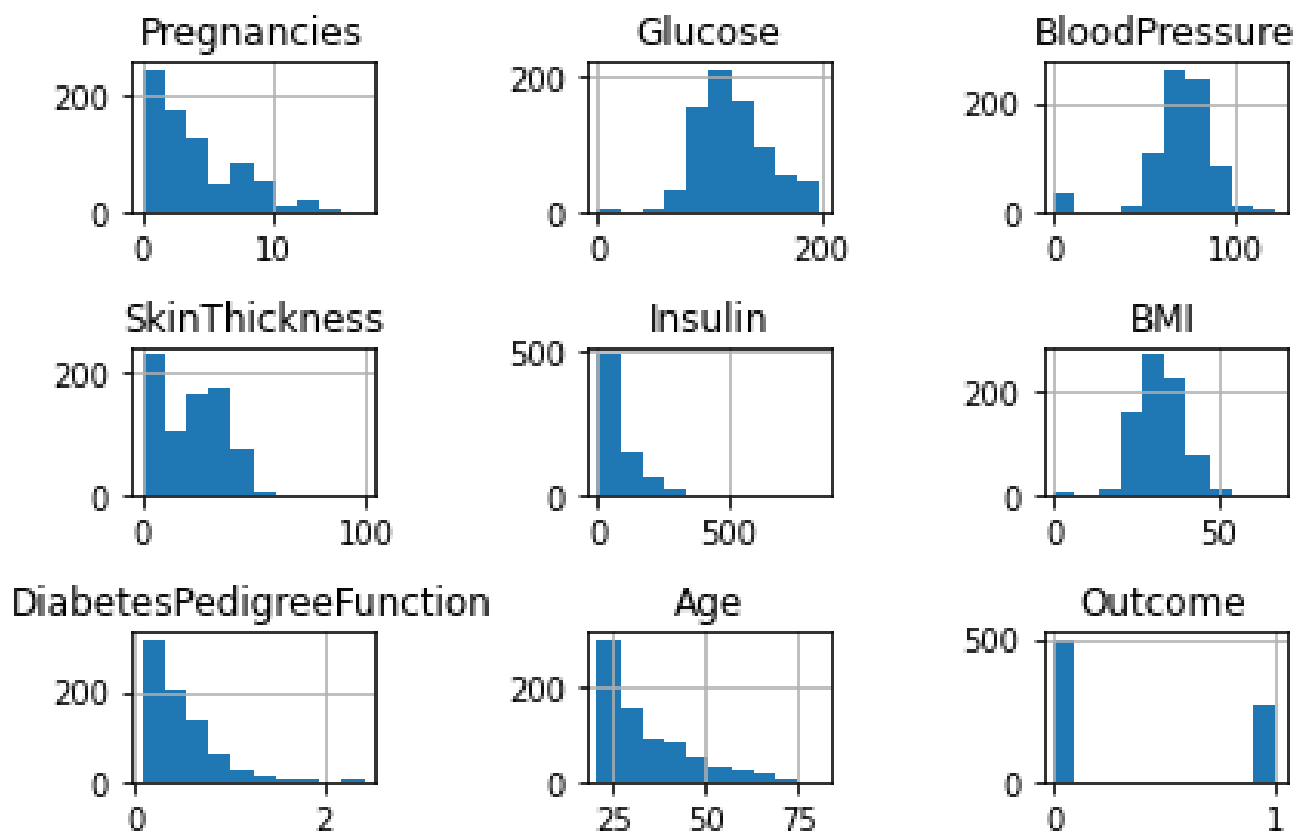
Figure 1: Correlation Analysis for All Factors Present in the Dataset.

As this heatmap indicates, 'Glucose', 'BMI' and 'Insulin' are top 3 variances impact on becoming type 2 diabetes. Since then, we would like to see how 'Glucose' potentially correlates the other two variances and relates to the diabetes.

Noise/Outliers



Pregnancies v.s Diabetes	Glucose v.s Diabetes	BloodPressure v.s Diabetes	SkinThickness c.s Diabetes
			
Insulin v.s Diabetes	BMI v.s Diabetes	DiabetesPedigreeFunction v.s Diabetes	Age v.s Diabetes



	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

```
[ ] df.skew()

Pregnancies      0.901674
Glucose           0.173754
BloodPressure    -1.843608
SkinThickness     0.109372
Insulin           2.272251
BMI               -0.428982
DiabetesPedigreeFunction  1.919911
Age               1.129597
Outcome           0.635017
dtype: float64
```

Based on visualization above, we see there are some outliers in those 8 variables. However, accounting to the data.describe(), we saw insulin has the greatest mean and variance, and all of them are skew distributed without null value.

ANOVA and Welch's T-test

Hypothesis

Null Hypothesis = Insulin and BMI correlates to glucose, if Insulin or BMI change, the glucose will change.

Alternative Hypothesis \neq Insulin and BMI correlates to glucose, if Insulin or BMI change, the glucose will not change.

I apply bin to group my individual sample of Glucose with two independent variables: BMI and Insulin. I also created bin group based on its median of Insulin and median of BMI.

bin1: low Insulin & low BMI;

bin 2 low Insulin & high BMI;

bin3: high Insulin & low BMI;

bin 4: high Insulin & high BMI

People with BMI < median are considered as underweight group

People with BMI > median are considered as overweight group

People with Insulin < median are considered as under-insulin group

People with Insulin > median are considered as over-insulin group

Data Cleaning for Linear Regression and ANOVA

1. Firstly, we want to see the distribution of Glucose value and what kind of degree it will potentially impact on diabetes. By cleaning the data, we got the median of glucose is 117 after flatten.

The Distribution of Glucose Values

df.describe()									
	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	0.471676	33.240885	0.348958
std	3.369578	31.972619	19.355907	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

	Glucose	Outcome
count	136.000000	136.000000
mean	128.808824	5.647059
std	41.416317	3.905418
min	0.000000	1.000000
25%	95.750000	3.000000
50%	129.500000	5.000000
75%	163.250000	8.000000
max	199.000000	17.000000

120.89453125

768

median is 117.0

Glucose

0	0
1	0
2	0
3	0
4	0

2. Then, we checked the BMI distribution, and got the median is 32 after flatten

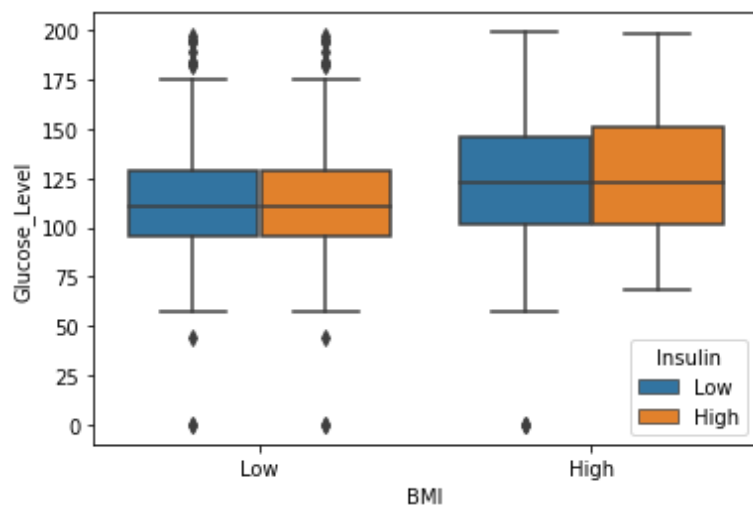
	BMI	Outcome	
count	248.000000	248.000000	31.992578124999977
mean	33.705645	3.096774	768
std	9.080065	2.359988	median is 32.0
min	0.000000	1.000000	
25%	26.575000	1.000000	BMI
50%	33.650000	2.000000	0 0.0
75%	39.825000	4.000000	1 0.0
max	67.100000	13.000000	2 0.0
			3 0.0
			4 0.0

3. Lastly, we checked the distribution of Insulin, and got the median is 30.5 after flatten

	Insulin	Outcome	
0	0	374	79.79947916666667
1	14	1	768
2	15	1	median is 30.5
3	16	1	
4	18	2	Insulin
			0 0
			1 0
			2 0
			3 0
			4 0

Based on the analysis above, we further generated the boxplot for those 4 bins to further understand how Glucose correlates with other two variances - Insulin, BMI. And we learnt that

- The glucose level in bin1 is not normally distributed
- The glucose level in bin2 is noramlly distributed
- The glucose level in bin3 is not normally distributed
- The glucose level in bin4 is not normally distributed



Summary and Discussion

OLS Regression Results

Dep. Variable:	Glucose_Level	R-squared:	0.035
Model:	OLS	Adj. R-squared:	0.032
Method:	Least Squares	F-statistic:	9.796
Date:	Tue, 29 Mar 2022	Prob (F-statistic):	2.37e-06
Time:	02:37:35	Log-Likelihood:	-3952.2
No. Observations:	811	AIC:	7912.
Df Residuals:	807	BIC:	7931.
Df Model:	3		

Covariance Type: nonrobust

	coef	std err	t	P> t	[0.025	0.975]
Intercept	127.3014	2.143	59.399	0.000	123.095	131.508
BMI[T.Low]	-13.1956	3.071	-4.297	0.000	-19.223	-7.168
Insulin[T.Low]	-2.8809	3.211	-0.897	0.370	-9.183	3.421
BMI[T.Low]:Insulin[T.Low]	2.8809	4.470	0.645	0.519	-5.893	11.655

Omnibus: 18.602 Durbin-Watson: 0.133
 Prob(Omnibus): 0.000 Jarque-Bera (JB): 36.524
 Skew: 0.056 Prob(JB): 1.17e-08
 Kurtosis: 4.034 Cond. No. 6.80

Warnings:

[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

	df	sum_sq	mean_sq	F	PR(>F)
BMI	1.0	28749.842015	28749.842015	28.581599	1.170204e-07
Insulin	1.0	392.042783	392.042783	0.389749	5.326080e-01
BMI:Insulin	1.0	417.837213	417.837213	0.415392	5.194285e-01
Residual	807.0	811750.342107	1005.886421	NaN	NaN

Above all, this project we created four thresholds which are based on the median values of the two explanatory variables "BMI" and "Insulin" as follow:

bin1: low Insulin & low BMI

bin 2: low Insulin & high BMI

bin3: high Insulin & low BMI

bin 4: high Insulin & high BMI

Meanwhile, we apply additional data transformations(encodings), ANOVA and OLS regression analysis, to see how those four bins contribute to the variation of glucose levels. And analysis

above helps inform patients to understand how it reduces their potential risks of having diabetes.

Our finding from the ANOVA and the subsidiary regression model are based on the explanatory variable BMI (with $PR(>F) = 0.00$). It is a significant factor contributing to the variation of one's glucose value, but the other variable Insulin level (with $PR(>F) = 0.07$) is not as much. That means the weight (body-mass index) or physical shape of a person influences on the certain degree of his glucose levels. And the patient's insulin does not influence on his measured glucose value. According to earlier research indicated, a healthier lifestyle can alleviate one's glucose level and potential risks of diabetes, so glucose value is moderately correlated with the diagnosis of type II diabetes.

For future iteration of project, we can improve this finding by building a more accurate regression model with a higher adjusted R value to further explore variations of the target variable.

Effect Size and Power Analysis

After conducting the ANOVA analysis in midterm, we further compared the effect sizes between the groups in the ANOVA to determine the real-world significance of the experiment.

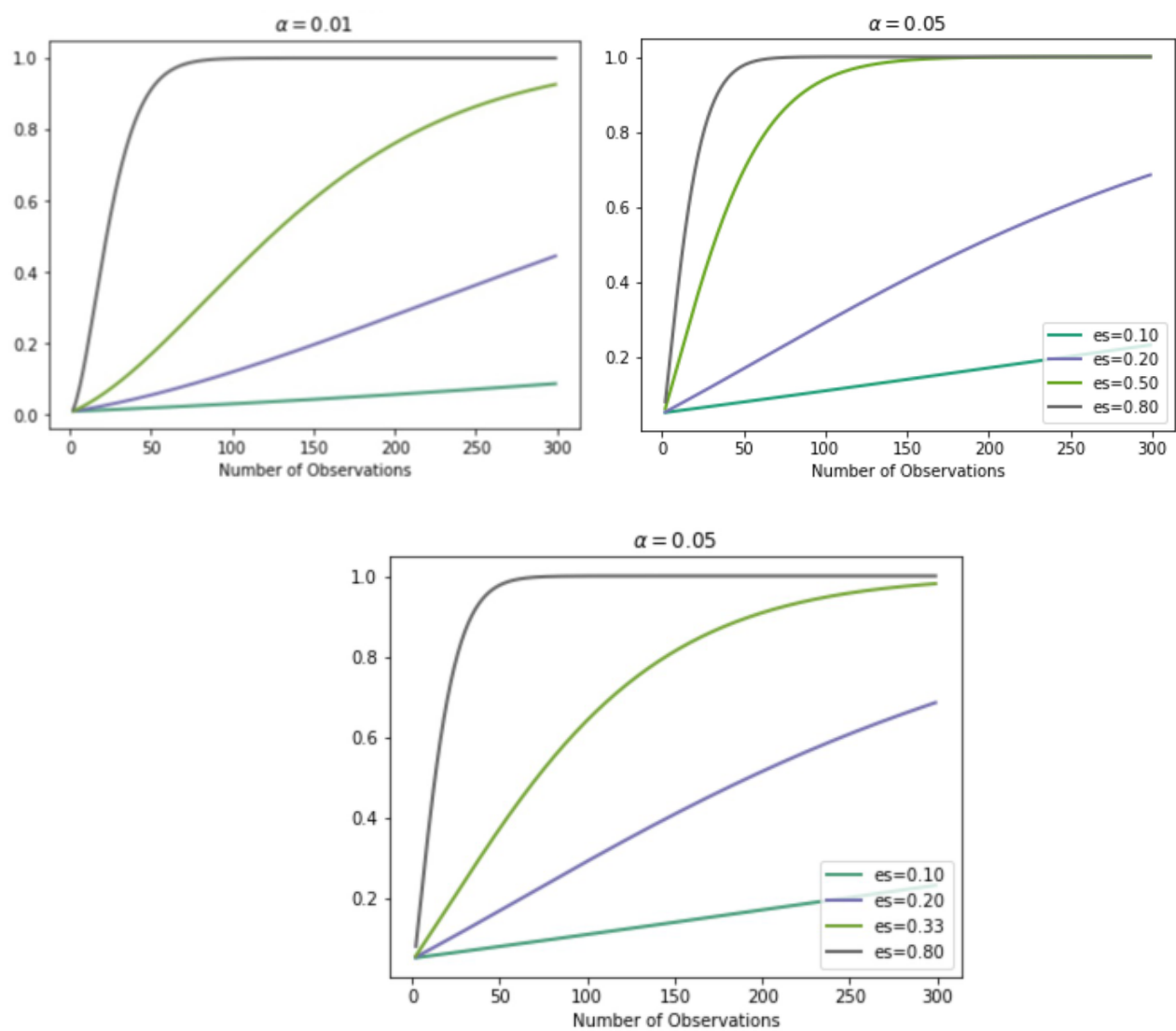
1. We first created group A including individuals with low BMI (combine bin1 and bin3), and group B for whom have high BMI (combine bin2 and bin4). We can see that the point estimate for the mean of glucose level measured from group A is 115.47 and group B is 126.02. Given the confidence level analysis, we find that 95% of all times, the mean value of glucose level of the low-BMI group is located between 112.47 and 118.46 mg/ml, and the high-BMI group is located between the interval 122.71 - 129.33 mg/ml.
2. Then, we created group C for people with low insulin levels and group D for whom with high insulin levels. Now, we can see that the point estimate for the mean of group C is 56.30 and 96.68 for group D.

So, the *confidence interval* for C and D are 112.47 - 118.46 mg/ml and 122.71 - 129.33mg/ml respectively.
3. The *effect size* (Cohen's d) between group A and B concerning difference in Glucose value, which impacted by BMI is 0.33 and 0.36 between group C and D. It suggests the differences revealed are small to moderate in real-world significance (or 60% of the glucose values in the higher group are higher than their counterparts in the lower group). The resultant

power analysis further indicates that, the required sample size for achieving an effect size of 0.33, statistical power is 0.8(type II error lower than 0.2) at a significance level of 0.05 is 145.11, suggesting that for each group sample there should be at least 146 observations.

The following images show that, when the effect size or "es" (difference between the group means calculated) is larger, it requires smaller sample sizes to achieve a higher statistical power.

Based on the power of t-Test,



ANCOVA and Linear Model with Interaction Model

	coef	std err	t	P> t	[0.025	0.975]
const		0.542	19.991	0.000	9.769	11.896

	10.8326					
BMI	0.9365	0.013	70.531	0.000	0.910	0.963
Insu	0.1106	0.003	35.109	0.000	0.104	0.117
interaction	-0.0052	0.000	-30.023	0.000	-0.006	-0.005

$$Y = 10.83 + 0.93 \cdot \text{BMI} + 0.11 \cdot \text{Insulin} - 0.0052 \cdot \text{BMI} \cdot \text{Insulin}$$

Adjusted R = 0.980; R = 0.980; F = 12390; P(F) = 0.00

This model explains 0.98 of the variations of the target variable which is accurate. With a significant interaction term, the main effects as the increasing BMI (the more a person becomes obese) correlated with the rising glucose value to different extents. It depends on the insulin level generated in his body. The negative coefficient of the insulin level suggests that a rising insulin level can alleviate the rising glucose value contributed by BMI (becoming obese), which is consistent with the current scientific finding that insulin helps lower one's glucose and the lack of insulin might result in type 2 diabetes (Healthline, no date).

Critical Discussion

The metric BMI is not considered as a robust measurement of the health impact on one's obesity, as it does not account for the difference between one's sex, bone structure and fat distribution. A muscular and athletic person might be inaccurately placed in the overweight and obese category, as the muscle of these athletes contributes to the weight factor of the BMI value. Also, it is revealed that one's waist size (fat accumulated there) is linked to diabetes risk regardless of one's BMI value. As a result, the BMI might be a fairly flawed predictor for one's diabetes risks (*Medical News Today, no.date*).

References

1. Tine D. Clausen; Elisabeth R. Mathiesen; Torben Hansen; Oluf Pedersen; Dorte M. Jensen; Jeannet Lauenborg; Peter Damm. (2008, Feb 1). High Prevalence of Type 2 Diabetes and Pre-Diabetes in Adult Offspring of Women With Gestational Diabetes Mellitus or Type 1 Diabetes: The role of intrauterine hyperglycemia. Retrieved from American Diabetes Association: <http://diabetesjournals-org.myaccess.library.utoronto.ca/care/article/31/2/340/25199/High-Prevalence-of-Type-2-Diabetes-and-Pre>

2. Julia Belluz. (2016, April 6). Why BMI is a flawed measure of body fat, explained by an eloquent 14-year-old: <https://www.vox.com/2016/4/6/11377158/bmi-flaws-tessa-embry>
3. Christian Nordqvist (2022, Jan 19). Why BMI is inaccurate and misleading: <https://www.medicalnewstoday.com/articles/265215#Waist-size-linked-to-diabetes-risk,-regardless-of-BMI>
4. Darren Hein; Pharm D; (2018, Dec 21). How Insulin and Glucagon Work: <https://www.healthline.com/health/diabetes/insulin-and-glucagon#:~:text=Insulin%20helps%20keep%20the%20glucose,%2C%20muscles%2C%20and%20fat%20tissue.>