Research Question

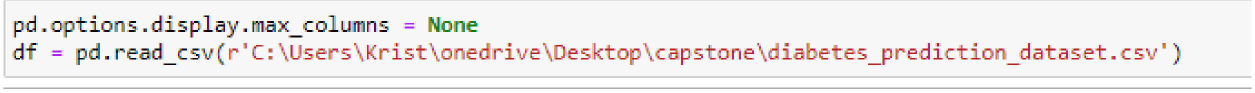
Can a linear regression model use BMI, HbA1C, age, and gender to predict presence or absence of heart disease? Heart disease can have various impacts on businesses, including healthcare costs, lost productivity, absenteeism, disability, and premature death of employees. Heart disease is a significant health issue in the United States, and it does have substantial economic implications. According to the American Heart Association (*AHA, A Costly Burden for America)*, cardiovascular diseases, including heart disease, will cost the United States an estimated $600 billion in direct healthcare expenses and lost productivity this year. Having a model to accurately predict the risk level of a person developing this disease can help businesses develop wellness interventions to keep employees healthy and working. CVD is the costliest disease and the top killer in the US, according to the same report. Businesses need an accurate model to determine which employees are at-risk of developing heart disease so that they can be targeted for wellness initiatives to reduce the productivity loss and increased medical costs affecting the company.

Data Collection

The data was found on Kaggle.com. Kaggle houses thousands of datasets, some preprocessed and cleaned, but many not. One advantage of using Kaggle to find the dataset is that there are so many sets available for the analyst to choose from. A disadvantage is that if the analyst wants to use certain variables that particular dataset may not be available. The analyst also has no control over the gathered information in terms of how the data was collected, and pre-processed. Due to the wide-range of subjects covered in the Kaggle datasets the only real challenge was narrowing the analyst's interests to one particular dataset. The chosen dataset contains 100,000 records of patient medical history. The records contain age, gender, BMI, HbA1C level, smoking history, hypertension history, diabetes history, heart disease history, and blood glucose level. There are no nulls or missing information.

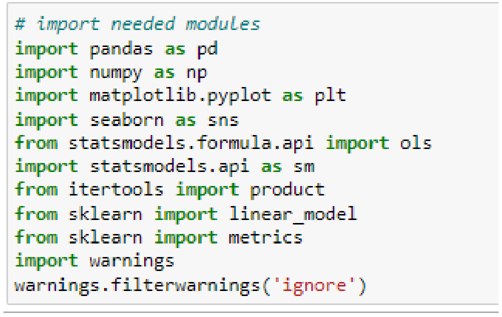
Data Extraction and Preparation

The dataset was downloaded from Kaggle and uploaded to the Jupyter Notebook environment, using Python. The data was read into a pandas dataframe with all columns shown.



```
pd.options.display.max_columns = None
df = pd.read_csv(r'C:\Users\Krist\onedrive\Desktop\capstone\diabetes_prediction_dataset.csv')
```

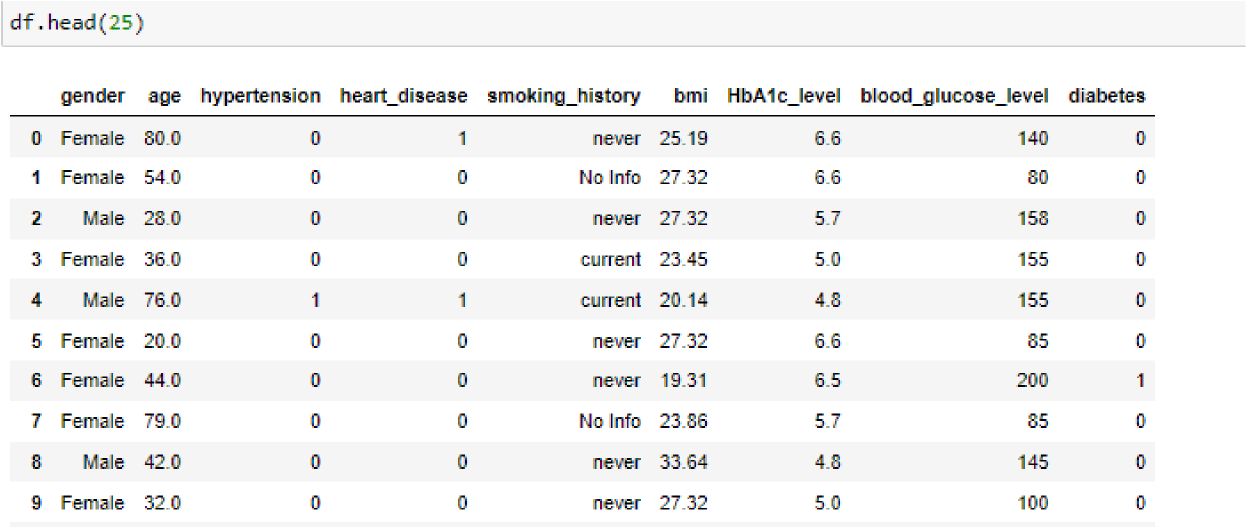
The usual Python libraries were also loaded, including pandas, numpy, matplotlib.pyplot, seaborn, and several sklearn modules specific to logistic regression. These libraries were each chosen for their specific capabilities. Pandas for its data manipulation allowing the dataset to be loaded into a DataFrame. The DataFrame allows for ease of changing data types, finding and replacing null values, group data, filter data, and perform simple statistics. Numpy is used for its



```python
# import needed modules
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
from statsmodels.formula.api import ols
import statsmodels.api as sm
from itertools import product
from sklearn import linear_model
from sklearn import metrics
import warnings
warnings.filterwarnings('ignore')
```

efficient handling of numerical arrays and matrices needed for the data modeling steps and mathematical calculations on the data. Matplotlib and seaborn allow the analyst to easily create visualizations such as bar charts, boxplots, scatter plots, histograms, and heatmaps. These visualizations allow the analyst to provide easy-to-read statistical analysis and informative graphs for the non-technical end user. Last, scikit-learn is a comprehensive machine learning library offering a wide range of modules and functions for data preprocessing and modeling. The data was scaled and normalized using scikit-learn.

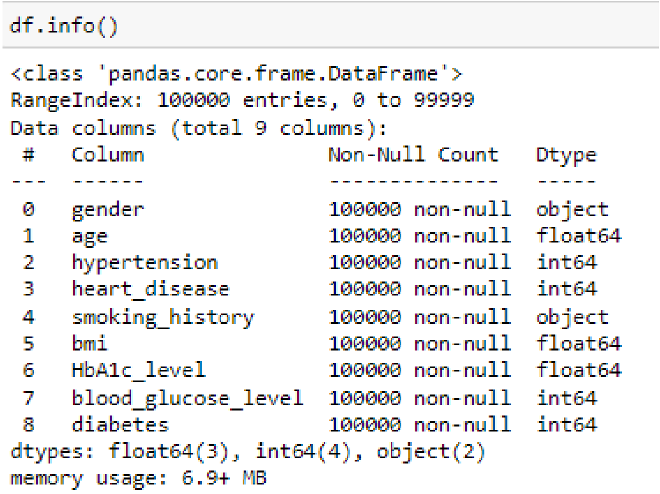
A quick check of the initial data shows no patient identifiers, including ID numbers, so all records can be referred to by their index position. Data is given for 100,000 records including: age, gender, hypertension, smoking history, bmi, HbA1c level, blood glucose level, and diabetes status. The data dictionary included with the dataset explained that hypertension, heart disease, and diabetes 0= no diagnosis in the patient's medical history, 1= positive patient medical history.



```
df.head(25)
```

	gender	age	hypertension	heart_disease	smoking_history	bmi	HbA1c_level	blood_glucose_level	diabetes
0	Female	80.0	0	1	never	25.19	6.6	140	0
1	Female	54.0	0	0	No Info	27.32	6.6	80	0
2	Male	28.0	0	0	never	27.32	5.7	158	0
3	Female	36.0	0	0	current	23.45	5.0	155	0
4	Male	76.0	1	1	current	20.14	4.8	155	0
5	Female	20.0	0	0	never	27.32	6.6	85	0
6	Female	44.0	0	0	never	19.31	6.5	200	1
7	Female	79.0	0	0	No Info	23.86	5.7	85	0
8	Male	42.0	0	0	never	33.64	4.8	145	0
9	Female	32.0	0	0	never	27.32	5.0	100	0

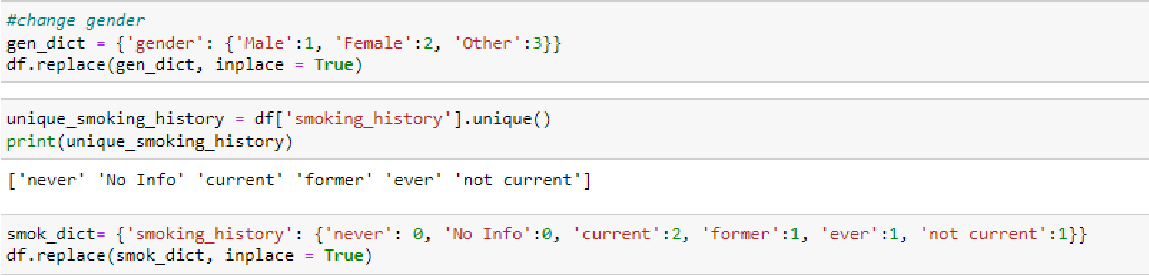
A logistic regression model requires all variables to be numerical. The info command shows two columns of type "object" that will need to be encoded to numerical data. Dictionaries were created to encode gender and smoking history. For smoking history, "no info" and "never" were combined, as were "former" and "not current". The analyst assumed that "no info" meant the patient had no history as a smoker, and "former" and "not current" meant the patient had a



```
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 100000 entries, 0 to 99999
Data columns (total 9 columns):
 #   Column               Non-Null Count   Dtype  
---  ------               --------------   -----  
 0   gender               100000 non-null  object 
 1   age                  100000 non-null  float64
 2   hypertension         100000 non-null  int64  
 3   heart_disease        100000 non-null  int64  
 4   smoking_history      100000 non-null  object 
 5   bmi                  100000 non-null  float64
 6   HbA1c_level          100000 non-null  float64
 7   blood_glucose_level  100000 non-null  int64  
 8   diabetes             100000 non-null  int64  
dtypes: float64(3), int64(4), object(2)
memory usage: 6.9+ MB
```

history of smoking but was not an active smoker. This assumption could put the analyst at a disadvantage by making incorrect assumptions. Descriptive statistics of the continuous variables give the analyst the beginnings of insight into the data gathered and how it may affect other variables.



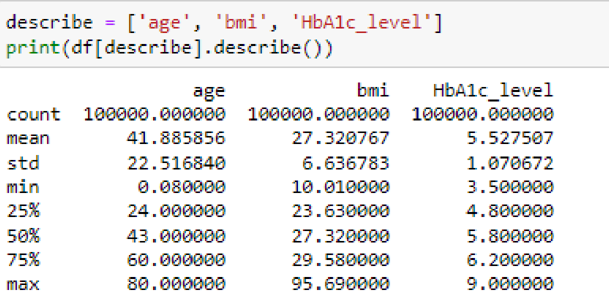
```
#change gender
gen_dict = {'gender': {'Male':1, 'Female':2, 'Other':3}}
df.replace(gen_dict, inplace = True)
```

```
unique_smoking_history = df['smoking_history'].unique()
print(unique_smoking_history)
```

```
['never' 'No Info' 'current' 'former' 'ever' 'not current']
```

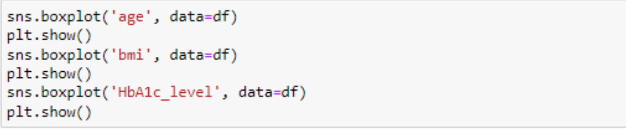
```
smok_dict= {'smoking_history': {'never': 0, 'No Info':0, 'current':2, 'former':1, 'ever':1, 'not current':1}}
df.replace(smok_dict, inplace = True)
```

The average age of patient is nearly 42, but patients represented in the data range from less than 1 year old to 80 years young. According to the American Cancer Society a normal BMI is 18.5-24.9, with the average patient here having a BMI in the overweight category, at 27.3. However the maximum BMI of 95.69 is far into the range of morbidly obese. A normal HbA1C level is less than 5.7, which the average patient is under that mark, meaning the last several months their body has been correctly regulating blood glucose levels. Seaborn boxplots were created for the continuous variables to find outliers. Boxplots are advantageous to use with continuous variables as it allows the analyst to easily spot outliers and determine if the data follows a normal distribution.

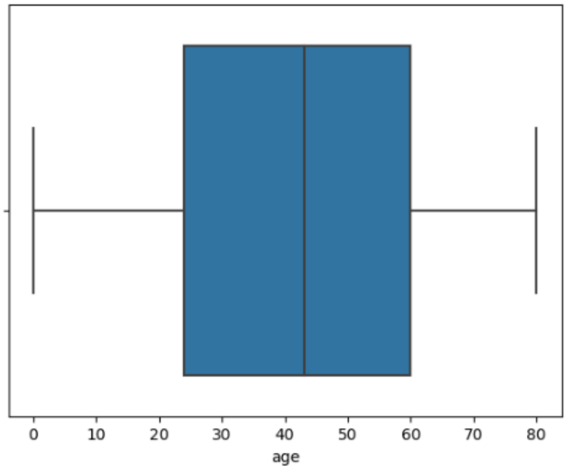


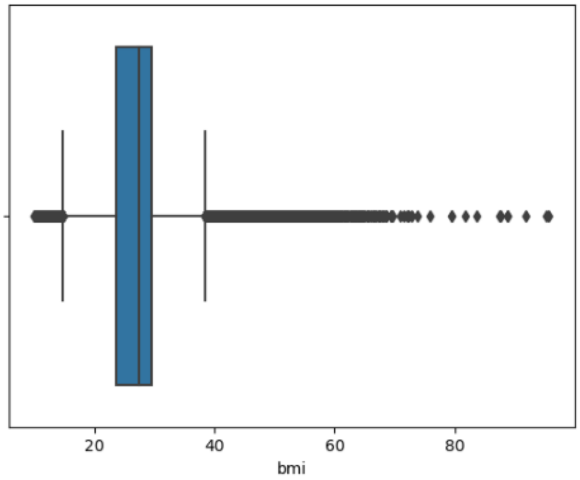
```
describe = ['age', 'bmi', 'HbA1c_level']
print(df[describe].describe())
```

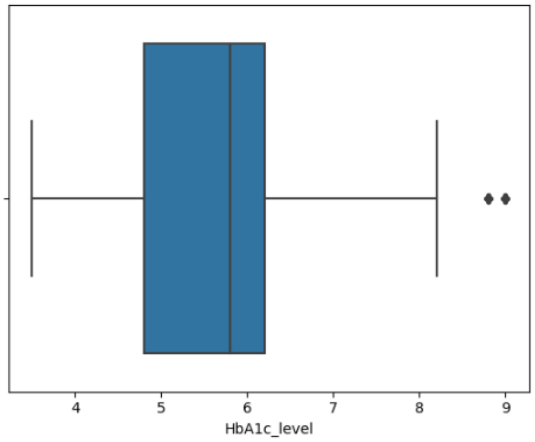
```
                 age            bmi    HbA1c_level
count  100000.000000  100000.000000  100000.000000
mean       41.885856      27.320767       5.527507
std        22.516840       6.636783       1.070672
min         0.080000      10.010000       3.500000
25%        24.000000      23.630000       4.800000
50%        43.000000      27.320000       5.800000
75%        60.000000      29.580000       6.200000
max        80.000000      95.690000       9.000000
```



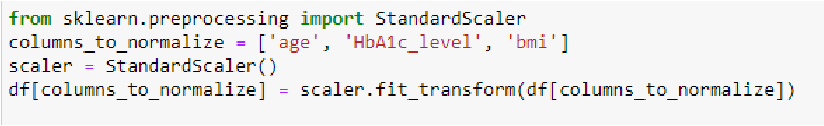
```
sns.boxplot('age', data=df)
plt.show()
sns.boxplot('bmi', data=df)
plt.show()
sns.boxplot('HbA1c_level', data=df)
plt.show()
```







As seen, the BMI data does not follow a normal distribution and has quite a few outliers specifically on the upper end. Trimming these outliers is not necessary in a logistic regression model and all data will remain in the algorithm. Lastly the data to be used in the model was normalized to ensure all data is on the same scale.
The data is now fully pre-processed and ready for the model to be created.



```
from sklearn.preprocessing import StandardScaler
columns_to_normalize = ['age', 'HbA1c_level', 'bmi']
scaler = StandardScaler()
df[columns_to_normalize] = scaler.fit_transform(df[columns_to_normalize])
```

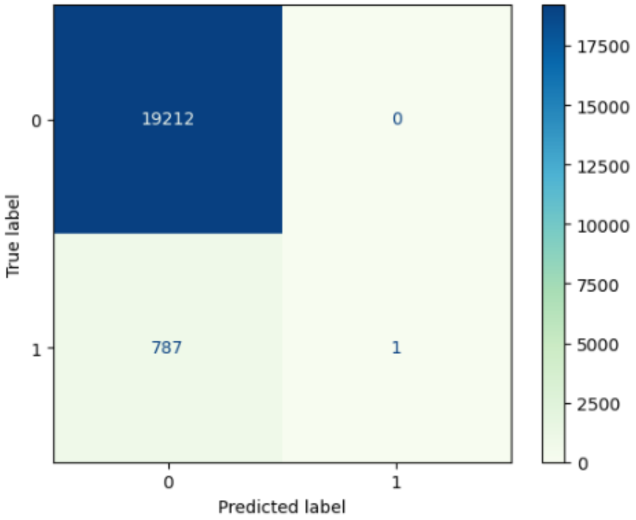
Analysis

Logistic regression provides interpretable results by estimating the impact of each feature on the probability of a particular outcome. The coefficients of logistic regression can be interpreted as the log-odds or the effect of the corresponding feature on the target variable. However, a disadvantage of logistic regression is that it assumes a linear relationship between the features and the log-odds of the target variable. This assumption may not hold in complex datasets where the relationship is non-linear. In such cases, logistic regression may not capture the underlying patterns accurately.

.Logistic regression was used to analyze the dataset and create a working model. All outliers were left in the data as patients with higher BMI and HbA1C levels are at a greater risk of heart disease, the target variable. After the model was created the features were ranked to determine if any could be dropped without negatively affecting the model. All features ranked #1 and therefore remained in the model. The model produced the following intercept and coefficients, resulting in the following equation: Probability of a heart disease diagnosis = -3.358* -.875 * gender * 1.81 * age * .363 * smoking history * .214 * BMI * .173 * HbA1C level. This shows men are at the greatest risk of developing heart disease and age is the most important of these five variables in predicting heart disease, the older the patient the higher the risk of heart disease being found. Of the five variables, HbA1C level has the least effect on predicting heart disease. Logistic regression was used because it can easily produce the probability of an event occurring, in this case, heart disease. Logistic regression is a quick model to create, fit, and train. However, it requires a very large number of records to produce an effective model that does not fall victim to overfitting.

Data Summary and Implications

The research question was investigated thoroughly and the null hypothesis rejected. Age, gender, BMI, HbA1C, and smoking history can be used to create a logistic regression model and accurately predict the presence, or absence, of heart disease. When tested against the trained model the model accurately predicted the presence or absence of heart disease 96% of the time. Less than 1% of patients that had heart disease were given an incorrect diagnosis. Due to the high costs to businesses due to employee heart disease, in both loss of productivity and medical costs incurred by the company, the model should be used to identify employees at risk of heart disease, specifically men with a history of smoking, and target those employees for wellness initiatives such as tobacco cessation programs, and ways to lower HbA1C through diet and exercise. The model is limited by the small amount of variables present, as it was created with only five dependent variables. Future study of the dataset should include how hypertension and heart disease are related and which gender is more at risk of developing hypertension based on overweight BMI.



Works Cited

American Heart Association. (n.d.). *Cardiovascular Disease: A Costly Burden for America*. Retrieved June 15, 2023, from https://www.heart.org/en/get-involved/advocate/federal-priorities/cardiovascular-disease-burden-report.

Normal weight ranges: Body mass index (BMI). Information and Resources about for Cancer: Breast, Colon, Lung, Prostate, Skin. (n.d.). https://www.cancer.org/cancer/risk-prevention/diet-physical-activity/body-weight-and-cancer-risk/adult-bmi.html

Dataset

Mustafa, M. (2023, April 8). *Diabetes prediction dataset*. Kaggle. https://www.kaggle.com/datasets/iammustafatz/diabetes-prediction-dataset