**Speaker Notes**

**Slide 2:**

* According to the CDC, heart disease is the leading cause of death in the United States and the number one killer in Canada, it is also the most costly disease in Canada, putting the greatest burden on our national health care system.
* A number of factors, individually or in combination, can lead to heart disease, namely: high blood pressure, high blood cholesterol, and smoking are key risk factors for heart disease. About 47% of people in the United States have at least one of these three risk factors. Several other medical conditions and lifestyle choices can also put people at a higher risk for heart disease, including: diabetes, obesity, an unhealthy diet, diets rich in saturated fat, physical inactivity, excessive alcohol use, stress and a family history of heart disease

**Slide 3:**

* Some of the risk factors include an unhealthy diet, physical inactivity, tobacco use and harmful use of alcohol
* These factors lead to obesity and raised blood pressure, blood glucose, blood lipids

**Slide 4:**

* The Framingham risk score test is used to estimate the 10-year cardiovascular risk of an individual, using data from about 11,600 patients
* Majority of the patients were at low risk, but there were 16% who were intermediate risk and 3% at high risk, with majority of the high risk being men

**Slide 5:**

The questions we hope to answer include:

* whether we can predict if certain individuals will develop coronary heart disease within the next 10 years given certain risk factors
* What is the correlation between different risk factors associated with heart disease
* How are the different risk factors correlated between males and females
* Which factors have more weight in predicting coronary heart disease

**Slide 6:**

The technolgies, languages, tools & alhorithms used:

* Machine Learning:
  + AWS connected to postgressql using sqlalchemy
  + Standardized using scikitlearn StandardScaler
  + RandomOverSampler from imblearn library
  + Scikit module test\_train\_split
* Data Cleaning & Analysis:
  + Pandas
  + Matplotlib
  + Seaborn
* Dashboard:
  + Tableau
* Model Choice:
  + Logistic Regression

**Slide 7:**

The Dataset Journey:

* Create AWS RDS
* Connect RDS to PostgreSQL
* Create temp and main tables
* Import csv files into temp tables
* Check data imported properly
* Clean and normalize data
* Copy clean data into main tables
* Join two tables, copy into a view
* Read view in Python, make a DF
* Remove records with null values
* Copy clean dataset into a table on AWS

**Slide 17:**

* We explored the column names which are the ones mentioned previously and the data types which are either float64 or int64
* We looked for null and duplicate values which we didn't find any

**Slide 18:**

* We then proceeded to explore graphically how the data in the different columns is distributed
* We performed basic statistic exploration with pandas .describe() method

**Slide 21:**

The first Machine Learning model we chose is Logistic Regression as we are trying to predict a discrete binary outcome. Logistic regression is easier to implement, interpret, and very efficient to train. It can have good accuracy for many simple data sets and it performs well when the dataset is linearly separable. Although we believe Logistic regression might work in this case, logistic regression inherently runs on a linear model and there are other models available which could also prove useful

**Slide 25:**

Some recommendations for a future analysis would be to acquire more data and to get more information on how important each feature is, to get a more accurate model

**Slide 26:**

The task that the team member took at the beginning is what they continued with throughout the project. Instead, we should have stuck more to this kind of work distribution, where the tasks are shuffled