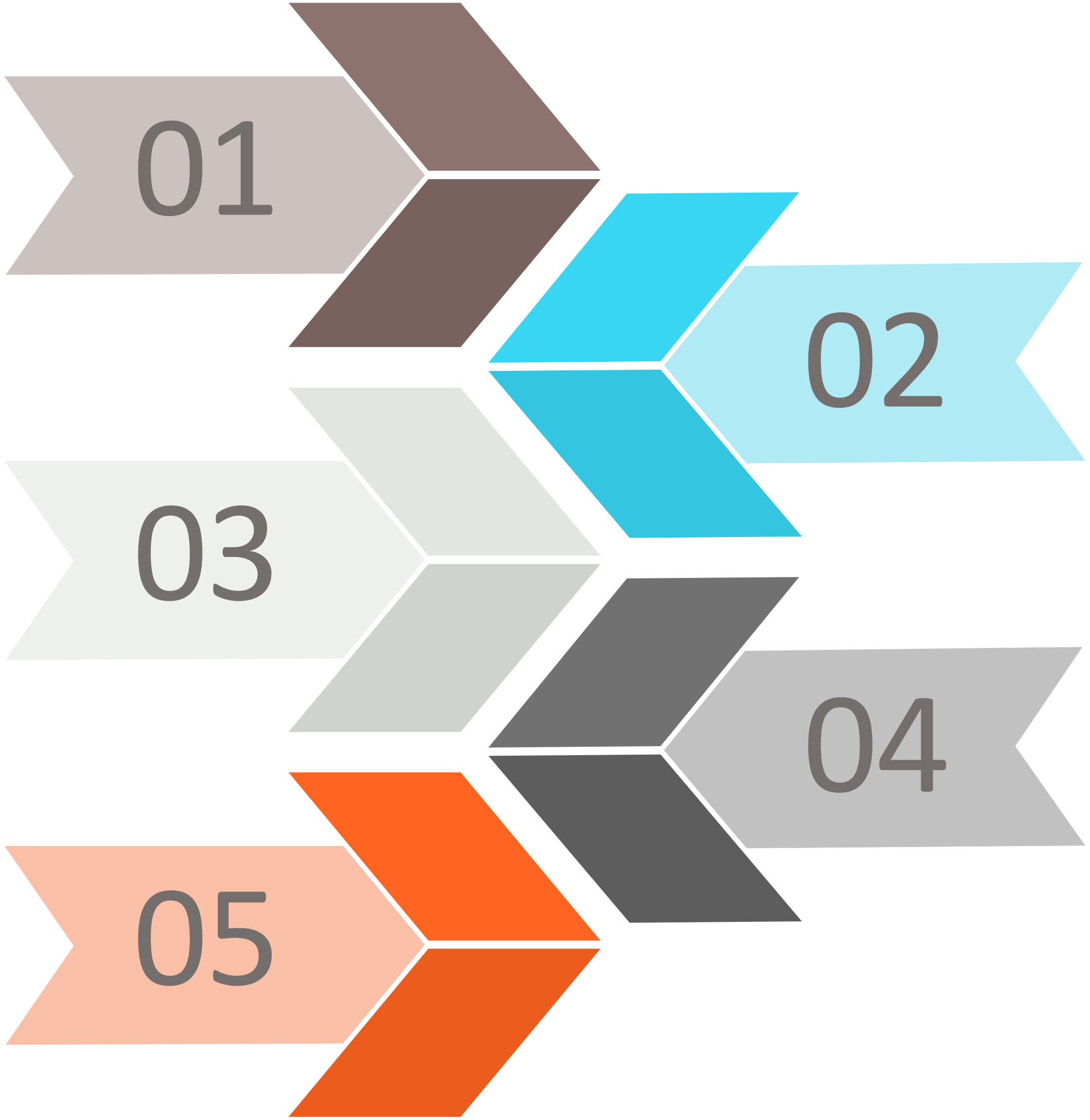
Deep Dive into Heart Disease Analysis

AGENDA



01 OBJECTIVE

02 BACKGROUND

03 KEY FINDINGS

04 RECOMMENDATIONS

05 APPENDIX

OBJECTIVE

1 Improve our shared understanding about the factors impacting heart disease

2 Improve our shared understanding regarding the insights found in heart disease predictions and anlysis

3 Provide early recommendations to the Analyst team

BACKGROUND

- The Average Age by Heart Disease : Those with target=1 (who have heart disease) have an average age of 52 years and those who have none appear to be in age group of 56-57 years
- The average age of the dataset is about 54 years and majority of them are males
- The mean for Chest Pain (cp)is closer to 1 (range is 0-3), suggesting that most patients have a low level of chest pain.
- Binary variable (0 for < 120 mg/dl, 1 for > 120 mg/dl). Only about 15% of patients have a fasting blood sugar above 120 mg/dl.
- Exercise Induced Angina (exang): Binary variable (0 for no, 1 for yes). About 34% of patients experience angina due to exercise.
- Number of Major Vessels (ca): Ranges from 0 to 4. The mean (0.75) suggests that most patients have fewer than one visible vessel on fluoroscopy.

INSIGHTS ON TARGET

- Distribution plots for 'age', 'chol', and 'thalach' based on the 'target' variable. The plots show that : 1. Individuals with heart disease (target = 1) tend to be peak around 50-55 age group. Those without heart disease (target = 0) tend to peak at 55-60 age group
- 2. Cholesterol (Chol): The distribution of cholesterol levels for individuals with heart disease appears slightly higher on average compared to those without heart disease. The peak for target=1, being 200-210 and peak for target =0 being 230-240.
- 3. Maximum Heart Rate (Thalach): Individuals with heart disease tend to have a higher maximum heart rate, with a peak around 170-180 beats per minute. The individuals with target = 0, show a blunt peak between 140-150, although those without heart disease generally exhibit lower maximum heart rates, with a more spread out distribution.
- The average Resting Blood Pressure (trestbps) and average Cholesterol Levels are higher for target =0. The both sexes, target=1 is higher than target =0. For all categories for chest pain, target=1 is higher than target=0, except in case of category 0
- **thal column vs target:** thal is highest for category 2 for both the target(heart disease) values (0 and 1) and lowest for category 0 for both the target values.

INSIGHTS ON AGE

- Relationship with age:: **1. Age vs. Cholesterol:** cholesterol levels might increase with age. **2. Age vs. Maximum Heart Rate (thalach):** maximum heart rate tends to decrease as age increases.
- restecg, age, sex: Overall, sex=1 generally exhibit slightly restecg across all age groups as compared to sex=0.
- Distribution of ca across age groups by sex: For the 30-39 and 40-49 age groups, both males and females have extremely low ca values, suggesting a lower prevalence of major vessels in these younger age groups. The highest ca values are observed in sex=1 aged 70-79
- Distribution of slope across age groups by sex: For all age groups, sex=1 tend to have a higher slope value compared to sex=0
- Distribution of oldpeak across age groups by sex: For both sex=1, the oldpeak value tends to increase with age, indicating a higher ST depression during exercise as people get older.
- For the age groups 30-39, 40-49, and 50-59, sex=0 generally have a upward trend. The highest oldpeak values are observed in the 70-79 age group for sex=1, the highest peak is in the age group of 60-69 for sex=0

INSIGHTS ON AGE & SEX

- cp, age and sex: The values are generally uniform for the younger age groups (40-69) for both sexes and lower for the older age groups (70-79) for both males and females. Sex=0 having higher values across all age groups, suggests that sex=0 tend to experience more chest pain or have a higher severity of chest pain compared to sex=1 of the same age group
- **Resting Blood Pressure (trestbps), age and sex:** Blood pressure tends to increase with age for sex=1, with the highest levels observed in the 70-79 age group.
- **Cholesterol (chol) vs age vs sex:** chol tends to increase with age for sex=1, with the highest levels observed in the 70-79 age group. Sex=0 is at its peak in the 60-69 age group. The spread of values is generally wider in the older age groups, indicating greater variability in cholesterol levels among older individuals.
- **Thalach - Maximum Heart Rate Achieved, age, sex:** For both sexes, the maximum heart rate achieved tends to decrease with age, which is expected as cardiovascular fitness and maximum heart rate typically decline with aging. Sex=1 has higher thalach levels as compared to sex=0.

INSIGHTS ON SEX

- When we compare mean, median and mode of age vs sex, the majority of age group is between 55-62 for sex=0 and 53-58 for sex=1
- **Age vs trestbps by Sex:** A slightly positive or upward trend is observed with her relationship between Age and trestbps, majority of the sex column being sex=1. It is most dense in 50-69 age group
- **Age vs thalach by sex:** A slightly negative or downward trend is observed with her relationship between Age and thalach, majority of the sex column being sex=1. It is most dense in 50-69 age group
- **cp (chest pain) column by sex :** cp is highest for category 0 for both the sexes and lowest for category 3 for both the sexes.
- **thal column by sex :** thal is highest for category 2 for both the sexes and lowest for category 0 for both the sexes.

INSIGHTS ON DATA MODELLING

- Logistic regression has been used along with stats model to understand the statistical summary of the features and dataset
- Used the VIF method for feature engineering
- Have drooped the column **thal and thalach** in the train and test data . Additionally, also drooped a third column for test data
- Have found the confusion matrix for test and train features and the ROC curve
- Top features for test are: **sex, oldpeak, cp, exang, ca, fbs**
- Top features for train are: **cp, restecg, exang, ca, fbs**
- One can conclude **ca, fbs and exang** are the top features for the target column
- The accuracy for train data is 82.4% whereas for test data it is 85%

RECOMMENDATIONS

- The age groups 50-60 and 70-79 should be observed as majority of the variables have high levels in these age groups
- The treatment of heart disease should be gender specific as the different heart disease variables differ across the sexes. Having said that it is important to note that the data is imbalanced with sex=1, being the majority in the dataset.
- Other factors such as diabetes, smoking, menopause , inflammatory diseases (arthritis and auto immune disorders) and other lifestyle factors should also be considered while examining the heart disease which are currently not included in the dataset.

APPENDIX - DATA METHODOLOGY

- There are no major multicollinearity or relationships observed between the variables
- The target column is balanced but the sex column is not.
- **Outlier treatment:** Cholesterol (chol) and ST Depression (oldpeak) show notable outliers, as indicated by points that lie far outside the upper whiskers. Other variables like Resting Blood Pressure (trestbps) and Maximum Heart Rate (thalach) also display some outliers, but they are closer to the main distribution.

- Capped the values at 450 for chol(cholesterol) column. Note: A human can have 450 mg cholesterol, but is requires urgent attention from doctors, For analysis and to include these cases, the value is capped at 450 mg.
- capped the value number of major vessels (0-3) - 'ca’ colored by fluoroscopy - as mentioned in the data dictionary. There is no value as 4, which is mentioned in the dataset
- Thal - 0 = normal; 1 = fixed defect; 2 = reversable defect - this is mentioned in the dictionary; hence the value is capped at 2 even though there was lot of outliers with the value 3. This could be an error, or a category not mentioned in the data dictionary.

- ***Attached doc for methodology*** - Microsoft Word Document