Obesity Level Predictor

1. Descriptive Statistics, Simple Exploration and Data Cleaning

In [ ]:
```
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import numpy as np
from sklearn.cluster import KMeans

# Reading the file
df = pd.read_csv('Users/examsman/Desktop/Projects/Capstone/ObesityDataSet.csv')

# Display the first 5 rows
df.head()
```

Out [ ]:

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II

In [ ]:
```
# Display descriptive statistics of each numerical attribute
df.describe()
```

Out [ ]:

	Age	Height	Weight	FCVC	NCP	CH2O	FAF	TUE
count	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000	2111.000000
mean	24.312600	1.701677	86.586058	2.419043	2.685628	2.008011	1.010298	0.657866
std	6.345968	0.093305	26.191172	0.533927	0.778039	0.612953	0.850592	0.608497
min	14.000000	1.450000	39.000000	1.000000	1.000000	1.000000	0.000000	0.000000
25%	19.947192	1.630000	65.473343	2.000000	2.658738	1.584812	0.124505	0.000000
50%	22.777890	1.700499	83.000000	2.385502	3.000000	2.000000	1.000000	0.625350
75%	26.000000	1.768464	107.430682	3.000000	3.000000	2.477420	1.666678	1.000000
max	61.000000	1.980000	173.000000	3.000000	4.000000	3.000000	3.000000	2.000000

In [ ]:
```
# Display the total null values in each attribute
df.isnull().sum()

# There are no missing values present in this dataset
```

Out [ ]:
```
Gender                            0
Age                               0
Height                            0
Weight                            0
family_history_with_overweight    0
FAVC                              0
FCVC                              0
NCP                               0
CAEC                              0
SMOKE                             0
CH2O                              0
SCC                               0
FAF                               0
TUE                               0
CALC                              0
MTRANS                            0
NObeyesdad                        0
dtype: int64
```

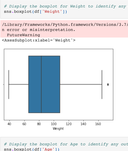
In [ ]:
```
# Number of total duplicated rows
print("Number of duplicated rows: ", df.duplicated().sum())
print("There are", df.duplicated().sum(), "duplicated rows" )

#Drop the duplicated rows
df = df.drop_duplicates()

# Re-count the total number of rows and display the type of data attributes
df.info()

# Total number of rows dropped from 2110 to 2087 rows
```

```
There are 24 duplicated rows
<class 'pandas.core.frame.DataFrame'>
Int64Index: 2087 entries, 0 to 2110
Data columns (total 17 columns):
 #   Column                          Non-Null Count  Dtype
---  ------                          --------------  -----
 0   Gender                          2087 non-null   object
 1   Age                             2087 non-null   float64
 2   Height                          2087 non-null   float64
 3   Weight                          2087 non-null   float64
 4   family_history_with_overweight  2087 non-null   object
 5   FAVC                            2087 non-null   object
 6   FCVC                            2087 non-null   float64
 7   NCP                             2087 non-null   float64
 8   CAEC                            2087 non-null   object
 9   SMOKE                           2087 non-null   object
 10  CH2O                            2087 non-null   float64
 11  SCC                             2087 non-null   object
 12  FAF                             2087 non-null   float64
 13  TUE                             2087 non-null   float64
 14  CALC                            2087 non-null   object
 15  MTRANS                          2087 non-null   object
 16  NObeyesdad                      2087 non-null   object
dtypes: float64(8), object(9)
memory usage: 293.5+ KB
```

In [ ]:
```
# Display the boxplot for Height to identify any outliers
sns.boxplot(df['Height'])
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  FutureWarning
<AxesSubplot:xlabel='Height'>
```

Out [ ]:



In [ ]:
```
# Display the boxplot for Weight to identify any outliers
sns.boxplot(df['Weight'])
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  FutureWarning
<AxesSubplot:xlabel='Weight'>
```

Out [ ]:

In [ ]:
```
# Display the boxplot for Age to identify any outliers
sns.boxplot(df['Age'])
```

```
/Library/Frameworks/Python.framework/Versions/3.7/lib/python3.7/site-packages/seaborn/_decorators.py:43: FutureWarning: Pass the following variable as a keyword arg: x. From version 0.12, the only valid positional argument will be `data`, and passing other arguments without an explicit keyword will result in a
n error or misinterpretation.
  FutureWarning
<AxesSubplot:xlabel='Age'>
```

Out [ ]:

In [ ]:
```
# Display the count of outliers for each attribute
Q1 = df.quantile(0.25)
Q3 = df.quantile(0.75)
IQR = Q3 - Q1
outliers = ((df < (Q1 - 1.5 * IQR)) | (df > (Q3 + 1.5 * IQR))).sum()

print(outliers)
```

```
Age                               167
Height                             27
CALC                                0
CH2O                                0
FAF                                 0
FAVC                                0
FCVC                                0
Gender                              0
Weight                              1
MTRANS                              0
NCP                               577
NObeyesdad                          0
SCC                                24
SMOKE                              44
TUE                                 0
CAEC                                0
family_history_with_overweight      0
dtype: int64
/Users/examsman/Desktop/Projects/Capstone/ObesityDataSet.csv:1: FutureWarning: Automatic reindexing on DataFrame vs Series comparisons is deprecated and will raise ValueError in a future version.  Do `left, right = left.align(right, axis=1, copy=False)` before e.g. `left == right`
```

We can see that NCP has a very high count of outliers. Since the attributes acts as an ordinal variable, we can keep it for now and remove the attribute if needed when developing the model. The most important attributes are Height and Weight for clustering, and we can see that it only has one outlier which shows that it does not need manipulation at the moment.

2. Deep Data Exploration and Preparation for Cluster Analysis

In [ ]:
```
# plotting correlation heatmap
sns.heatmap(df.corr(),annot=True, cmap='coolwarm')

# Displaying heatmap
plt.show()
```

Out [ ]:

The correlation heatmap indicates that none of the variables have a significant correlation to each other that needs to be dealt with. Height and Weight have the strongest positive correlation of 46%. Since we know both variables are crucial in determining the Obesity Level, it will remain in the dataset.

In [ ]:
```
# Create a copy of the original dataframe for data exploration
df_explore = df.copy()
# Calculate BMI using the original formula and Weight and Height columns
df_explore['BMI'] = df_explore['Weight'] / df_explore['Height']**2
# Display first 5 rows
df_explore.head()
```

Out [ ]:

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad	BMI
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight	24.386526
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight	24.238227
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight	23.765432
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I	26.851852
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II	28.342381

In [ ]:
```
# Apply logic to create another column named "Trick" to display if the BMI level falls into the correct category.
df_explore['Trick'] = np.where((df_explore['BMI'] < 18.5) & (df_explore['NObeyesdad'] == 'Insufficient_Weight'), 1
|(df_explore['BMI'] >= 18.5) & (df_explore['BMI'] < 25) & (df_explore['NObeyesdad'] == 'Normal_Weight'), 1
|(df_explore['BMI'] >= 25) & (df_explore['BMI'] < 30) & ((df_explore['NObeyesdad'] == 'Overweight_Level_I')|(df_explore['NObeyesdad'] == 'Overweight_Level_II')), 1
|(df_explore['BMI'] >= 30) & (df_explore['NObeyesdad'] == 'Obesity_Type_I'), 1
|(df_explore['BMI'] >= 35) & (df_explore['NObeyesdad'] == 'Obesity_Type_II'), 1
|(df_explore['BMI'] >= 40) & (df_explore['NObeyesdad'] == 'Obesity_Type_III'), 1, False)
# Calculate the accuracy of BMI when compared to the class label.
bmi_acc = df_explore['Trick'].sum() / len(df_explore)
print('The Accuracy of BMI and NObeyesdad is', bmi_acc)
```

```
The Accuracy of BMI and NObeyesdad is 95.11260182079539
```

This shows that BMI is not directly correlated to the Obesity level since not all levels are True. This means that some of the other column come into the equation into determining the proper diagnosis for Obesity Level. The class label is not completely determined by the BMI (Height vs Weight) but have a very accurate representation.

In [ ]:
```
# Break the Data into Gender and Class Label to make sure there is accurate representation
df.groupby(['NObeyesdad', 'Gender']).size().unstack().sort_values(['Female', 'Male'], ascending=False).plot(kind='bar')
```

Out [ ]:
```
<AxesSubplot:xlabel='NObeyesdad'>
```


As we can see, most of the data is accurately represented for each gender except for Obesity_Type_I and II. Both categories lack severe representation of opposite genders. This could be an issue if other factors such as body fat percentage and muscle mass into play since it can distort the class label. On the other hand, we do have Height and Weight attributes that can contribute into lowering the errors of the other factors.

In [ ]:
```
# Plot Height vs. Height to prepare for Clustering. Having a visual can show us if there is any cluster clearly depicted.
plt.scatter(df_explore['Height'], df_explore['Height'])
plt.xlabel('Height (m)')
plt.ylabel('Height (kg)')
plt.title('Height vs. Height')
# Display the plot
plt.plot(np.unique(df_explore['Height']), np.poly1d(np.polyfit(df_explore['Height'], df_explore['Height'], 1))(np.unique(df_explore['Height'])), color = 'black')
```

Out [ ]:
```
[<matplotlib.lines.Line2D at 0x7ffa84c3c110>]
```


Completing the Cluster Analysis

In [ ]:
```
# How to check the right number of clusters for annual income vs spending scores:
X = df.iloc[:,[1,2]].values
wcss = []

for i in range(1,11):
    km = KMeans(n_clusters = i, init = 'k-means++', max_iter=300, n_init = 10, random_state=0)
    km.fit(X)
    wcss.append(km.inertia_)

plt.plot(range(1,11), wcss)
plt.xlabel('k value')
plt.ylabel('wcss')
plt.title('Elbow Analysis')
plt.show()

# Best number of clusters is 4 as the curve smoothens after 4.
# Keep in mind that this is only for height vs weight
```

Out [ ]:

According to the elbow method, the best values of k is 4 clusters as the curve smoothens. However, we will still use 7 since we want to cluster based on each Obesity Level. On a more realistic note, having only 4 clusters that represent Underweight, Normal Weight, Overweight, and Obese would be the most effective way to develop the most accurate model based on clustering.

In [ ]:
```
# Number of Clusters:
K = 7

km = KMeans(n_clusters = K, init = 'k-means++', max_iter=300, n_init = 10, random_state=0)
y_means = km.fit_predict(X)

# Select random observations as centroids
fig=plt.figure(figsize=(6,6))
plt.scatter(X[y_means == 5, 1], X[y_means == 5, 1], s = 100, c = 'red', label = 'Insufficient_Weight')
plt.scatter(X[y_means == 0, 1], X[y_means == 0, 1], s = 100, c = 'blue', label = 'Normal_Weight')
plt.scatter(X[y_means == 1, 1], X[y_means == 1, 1], s = 100, c = 'cyan', label = 'Overweight_Level_I')
plt.scatter(X[y_means == 6, 1], X[y_means == 6, 1], s = 100, c = 'grey', label = 'Overweight_Level_II')
plt.scatter(X[y_means == 2, 1], X[y_means == 2, 1], s = 100, c = 'yellow', label = 'Obesity_Type_I')
plt.scatter(X[y_means == 4, 1], X[y_means == 4, 1], s = 100, c = 'orange', label = 'Obesity_Type_II')
plt.scatter(X[y_means == 3, 1], X[y_means == 3, 1], s = 100, c = 'green', label = 'Obesity_Type_III')
plt.scatter(km.cluster_centers_[:,1], km.cluster_centers_[:, 1], s = 50, c = 'blue' , label = 'Centeroid')
plt.xlabel('Height (m)')
plt.ylabel('Weight (kg)')
plt.legend()
plt.grid()
plt.show()
```

Out [ ]:

Comparing original Class Label and Clustering to measure Performance

In [ ]:
```
# Export the clustered results to a columns
df['Cluster'] = y_means
df.head()
```

Out [ ]:

	Gender	Age	Height	Weight	family_history_with_overweight	FAVC	FCVC	NCP	CAEC	SMOKE	CH2O	SCC	FAF	TUE	CALC	MTRANS	NObeyesdad	Cluster
0	Female	21.0	1.62	64.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	0.0	1.0	no	Public_Transportation	Normal_Weight	0
1	Female	21.0	1.52	56.0	yes	no	3.0	3.0	Sometimes	yes	3.0	yes	3.0	0.0	Sometimes	Public_Transportation	Normal_Weight	5
2	Male	23.0	1.80	77.0	yes	no	2.0	3.0	Sometimes	no	2.0	no	2.0	1.0	Frequently	Public_Transportation	Normal_Weight	2
3	Male	27.0	1.80	87.0	no	no	3.0	3.0	Sometimes	no	2.0	no	2.0	0.0	Frequently	Walking	Overweight_Level_I	6
4	Male	22.0	1.78	89.8	no	no	2.0	1.0	Sometimes	no	2.0	no	0.0	0.0	Sometimes	Public_Transportation	Overweight_Level_II	4

In [ ]:
```
# Apply logic to create another column named "Trick" to display if the Cluster falls into the correct category.
df['Trick'] = np.where((df['Cluster'] == 5) & (df['NObeyesdad'] == 'Insufficient_Weight'), 1
|(df['Cluster'] == 1) & (df['NObeyesdad'] == 'Normal_Weight'), 1
|(df['Cluster'] == 0) & (df['NObeyesdad'] == 'Overweight_Level_I'), 1
|(df['Cluster'] == 6) & (df['NObeyesdad'] == 'Overweight_Level_II'), 1
|(df['Cluster'] == 4) & (df['NObeyesdad'] == 'Obesity_Type_I'), 1
|(df['Cluster'] == 3) & (df['NObeyesdad'] == 'Obesity_Type_II'), 1
|(df['Cluster'] == 2) & (df['NObeyesdad'] == 'Obesity_Type_III'), True, False)
# Calculate the accuracy of Cluster when compared to the class label.
km_acc = df['Trick'].sum() / len(df)*100
print('The Accuracy of Cluster when compared to original class label is', km_acc*0.01, '%')
```

```
The Accuracy of Cluster when compared to original class label is 0.2994192127929 %
```

When comparing the cluster to the original class label, we can see that the K-means cluster analysis did not do a good job predicting the level of obesity correctly using 7 clusters. This is perhaps the amount of clusters are too much, or the height and weight were not the best indicators to diagnose obesity. In the Final results submission, I will include a code that runs the K-means Algorithm with just 4 clusters to see if it can develop a more accurate class label. For example, the 4 clusters can be Underweight, Normal, Overweight, and Obese. Since the elbow method supported this method, it will be interesting to see how well it will do.

Data Preparation for Supervised Learning (To be continued)

In [ ]:
```
num_col = df_clean._get_numeric_data().columns.tolist()

cat_col = df_clean.select_dtypes(include='object').columns.tolist()
```