Exploring BRFSS data

This report summarizes the findings from analyzing a dataset on weight and height collected by the Behavioral Risk Factor Surveillance System (BRFSS) survey. The BRFSS is a large-scale phone survey that gathers health data from US residents. The dataset used in this analysis contains six valuable columns: age, weight2 (current weight in kg), wtyrago (weight year ago in kg), wtkg2 (weight in 2 decimal places), htm3 (height in cm), sex (for males, the value is 1 and for females, it is 2), and NaN demonstrates the null values.

The data structure looks like the following picture, where the unnamed one is just the serial number of the number of entries.

	Unnamed: 0	age	weight2	wtyrago	wtkg2	htm3	sex
0	0	39.0	88.636364	88.636364	88.64	180.0	1
1	1	64.0	75.000000	84.545455	75.00	155.0	2
2	2	87.0	61.818182	63.636364	61.82	NaN	2
3	3	51.0	100.000000	100.000000	100.00	183.0	1
4	4	35.0	63.636364	61.363636	63.64	170.0	2

The data is also checked for null values.

```
# view the details of the dataset to check if data contains null values
assessment3.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414509 entries, 0 to 414508
Data columns (total 7 columns):
 #   Column      Non-Null Count   Dtype
---  ------      --------------   -----
 0   Unnamed: 0  414509 non-null  int64
 1   age         410856 non-null  float64
 2   weight2     398484 non-null  float64
 3   wtyrago     390399 non-null  float64
 4   wtkg2       398484 non-null  float64
 5   htm3        409129 non-null  float64
 6   sex         414509 non-null  int64
dtypes: float64(5), int64(2)
memory usage: 22.1 MB
```

After that, the unwanted unnamed column is removed, and statistical values are calculated for further preprocessing.

	age	weight2	wtyrago	wtkg2	htm3	sex
count	410856.000000	398484.000000	390399.000000	398484.000000	409129.000000	414509.000000
mean	54.862180	78.992337	79.721319	78.992453	168.825190	1.624368
std	16.737702	19.546212	20.565164	19.546157	10.352653	0.484286
min	18.000000	20.000000	22.727273	20.000000	61.000000	1.000000
25%	43.000000	64.545455	64.545455	64.550000	160.000000	1.000000
50%	55.000000	77.272727	77.272727	77.270000	168.000000	2.000000
75%	67.000000	90.909091	90.909091	90.910000	175.000000	2.000000
max	99.000000	309.090909	342.272727	309.090000	236.000000	2.000000

From the statistical table, the median is chosen for the preprocessing data filling, as filling is always better than removing it completely. The picture below shows that no null values are there now.

```
#Now after preprocessing view the changes
assessment3.info()
```

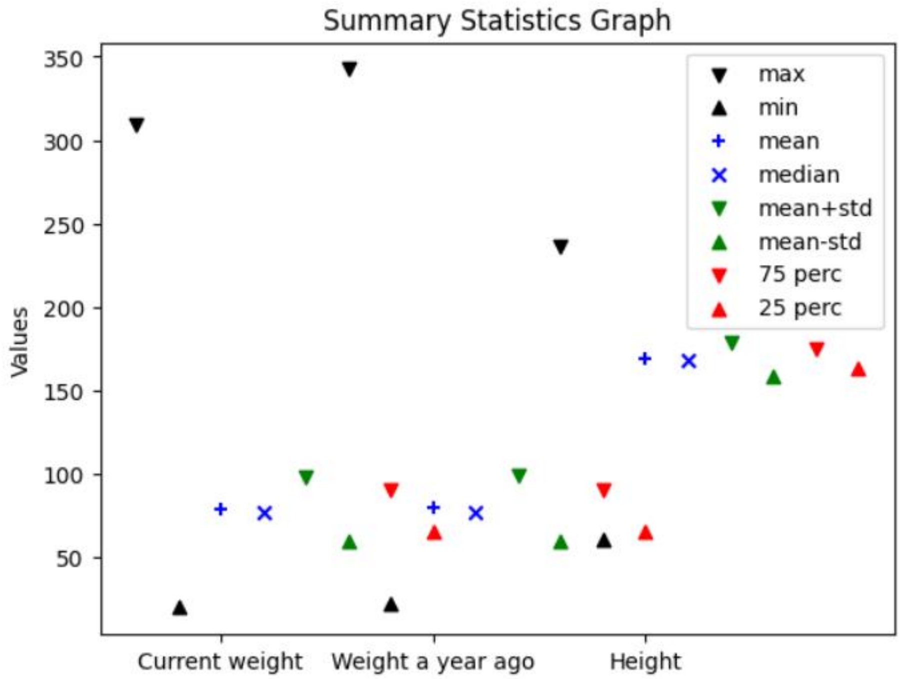
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 414509 entries, 0 to 414508
Data columns (total 7 columns):
 #   Column       Non-Null Count   Dtype
---  ------       --------------   -----
 0   Unnamed: 0   414509 non-null  int64
 1   age          414509 non-null  float64
 2   weight2      414509 non-null  float64
 3   wtyrago      414509 non-null  float64
 4   wtkg2        414509 non-null  float64
 5   htm3         414509 non-null  float64
 6   sex          414509 non-null  int64
dtypes: float64(5), int64(2)
memory usage: 22.1 MB
```

Since there are no null values and the data is preprocessed, Task 1 is initiated.

Section 1: Summary statistics analysis

The code calculates various summary statistics for weight (current and a year ago) and height. These statistics include mean, median, standard deviation, quartiles (25th and 75th percentile), minimum, and maximum values.

A visualization is generated to compare these statistics across the three variables. The key observation from this analysis is that the distribution of weight (current and a year ago) is likely right-skewed, as indicated by the higher values for the mean compared to the median. In contrast, the distribution of height appears to be closer to normal, with the mean and median values being relatively similar. Also, there are some outliers among weights, as demonstrated by the following picture.



Section 2: Correlations analysis

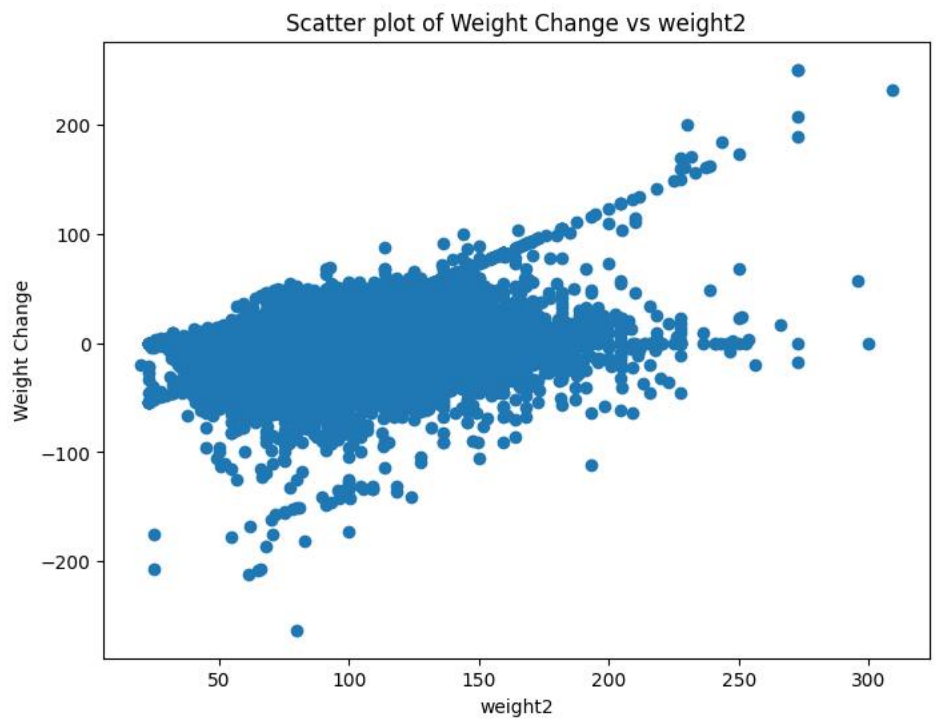
The code calculates the correlation coefficient between weight change (current weight minus weight a year ago) and three factors: weight2 (current weight), wtyrago (weight a year ago), and age. The correlation measures the strength and direction of the linear relationship between two variables. A correlation coefficient of 1 indicates a perfect positive linear relationship, -1 indicates a perfect negative linear relationship, and 0 indicates no linear relationship. Scatter plots are generated to visualize the relationships between weight change and each of the three factors. These plots confirm the findings from the correlation coefficients described above.

The analysis reveals the correlation of weight change with weight2, wtyrago, and age.

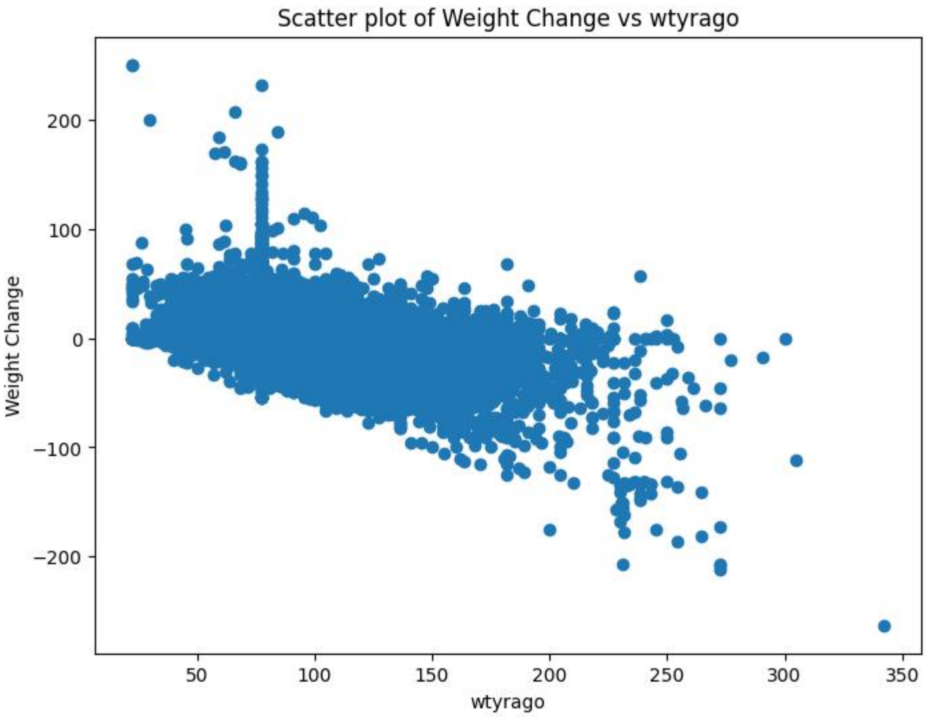
```
weight_change    1.000000
weight2          0.093601
wtyrago         -0.294092
age             -0.072108
Name: weight_change, dtype: float64
```

weight_change (1.000000): This is a perfect positive correlation, which means weight change has a perfect positive correlation with itself.

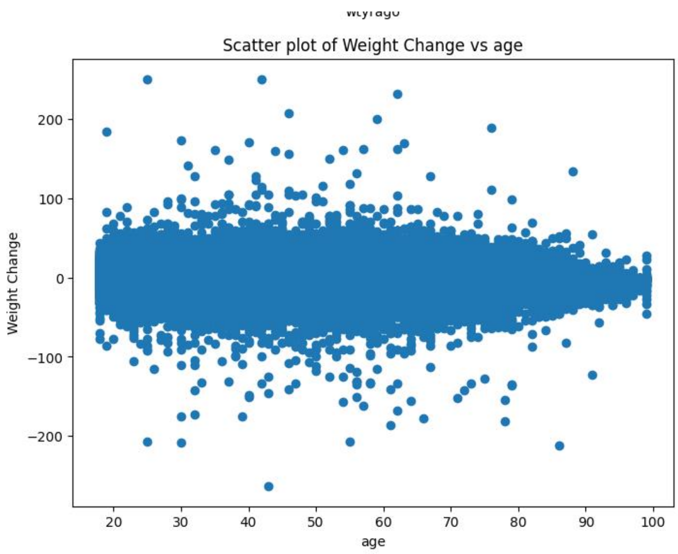
weight2 (0.093601): This is a weak positive correlation. It implies that people with a higher current weight (weight2) tend to have a larger weight change (either increase or decrease) compared to those with a lower current weight.



wtyrago (-0.294092): This is a weak negative correlation. It suggests that people with a higher weight a year ago (wtyrago) tend to have a smaller weight change (weight loss or minimal gain) compared to those with a lower weight a year ago. This is somewhat expected because current weight was influenced by weight a year ago.



age (-0.072108): This is a very weak negative correlation, close to zero. It indicates practically no linear relationship between weight change and age within this dataset.



Section 3: Linear regression analysis

This section demonstrates a simple linear regression model using scikit-learn. The model is trained on a dataset of house sizes and their corresponding prices to predict the price of a house with a specific area (2500 sq ft). Linear regression assumes a linear relationship between the independent variable (house size) and the dependent variable (house price).

The code successfully trains the model and predicts the price for a 2500-square-foot house. This demonstrates the application of linear regression for making predictions based on historical data. The final predicted price for a 2500-square-foot area is 326000.0.

```python
# linear regression model using scikit-learn, for training it on given house sizes and prices and predicts the price of a 2500 sqft house.
from sklearn.linear_model import LinearRegression
house_sizes = [[1500], [2000], [2500], [3000], [3500]]
house_prices = [250000, 300000, 330000, 360000, 390000]
model = LinearRegression()
model.fit(house_sizes, house_prices)
predicted_price = model.predict([[2500]])
print(f"For 2500 sqft area predicted house price is: {predicted_price[0]}")
```

```
For 2500 sqft area predicted house price is: 326000.0
```

Section 4: Conclusion

The analysis of the BRFSS data provides insights into weight change patterns. The findings suggest that current weight is the most influential factor associated with weight change. Additionally, there is no significant correlation between weight change and age within the scope of this data. The demonstration of linear regression highlights a technique for making predictions based on trends in existing data. Further analysis could explore gender differences in weight change patterns or investigate the relationship between weight change and other health factors included in the BRFSS survey.