Hyperiondev

TASK

# Exploratory Data Analysis on the food choice Dataset

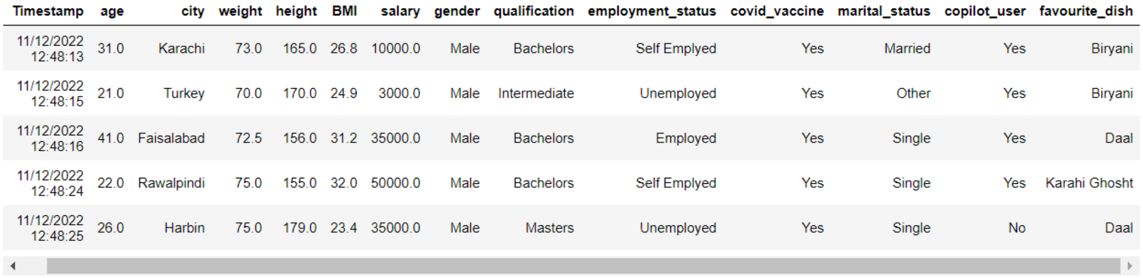
Visit our website

# Introduction

This dataset was obtained from Kaggle.com. The dataset shows food preferences of males and females with respect to their education, salary, BMI, City and employment status.

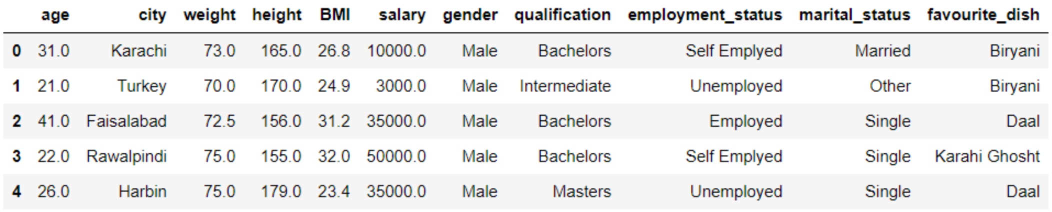
## DATA CLEANING

The food choice data was read into Jupyter notebook using a Pandas method. The data was looked at to get an idea of what was contained in the data frame. This was done, using the '.head()' method on the data.



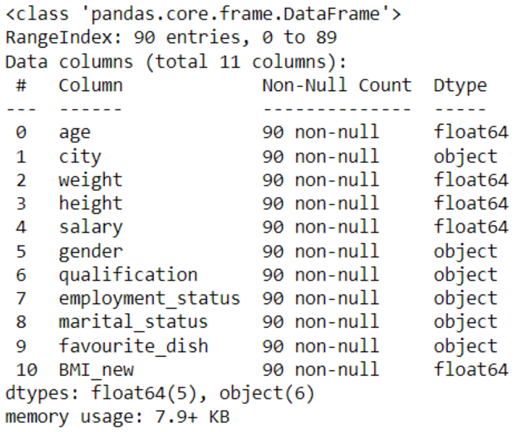
| Timestamp | age | city | weight | height | BMI | salary | gender | qualification | employment_status | covid_vaccine | marital_status | copilot_user | favourite_dish |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 11/12/2022 12:48:13 | 31.0 | Karachi | 73.0 | 165.0 | 26.8 | 10000.0 | Male | Bachelors | Self Emplyed | Yes | Married | Yes | Biryani |
| 11/12/2022 12:48:15 | 21.0 | Turkey | 70.0 | 170.0 | 24.9 | 3000.0 | Male | Intermediate | Unemployed | Yes | Other | Yes | Biryani |
| 11/12/2022 12:48:16 | 41.0 | Faisalabad | 72.5 | 156.0 | 31.2 | 35000.0 | Male | Bachelors | Employed | Yes | Single | Yes | Daal |
| 11/12/2022 12:48:24 | 22.0 | Rawalpindi | 75.0 | 155.0 | 32.0 | 50000.0 | Male | Bachelors | Self Emplyed | Yes | Single | Yes | Karahi Ghosht |
| 11/12/2022 12:48:25 | 26.0 | Harbin | 75.0 | 179.0 | 23.4 | 35000.0 | Male | Masters | Unemployed | Yes | Single | No | Daal |

Columns that were not necessary or redundant were removed. These columns include: 'Timestamp', 'copilot_user', 'covid_vaccine' and 'BMI'.



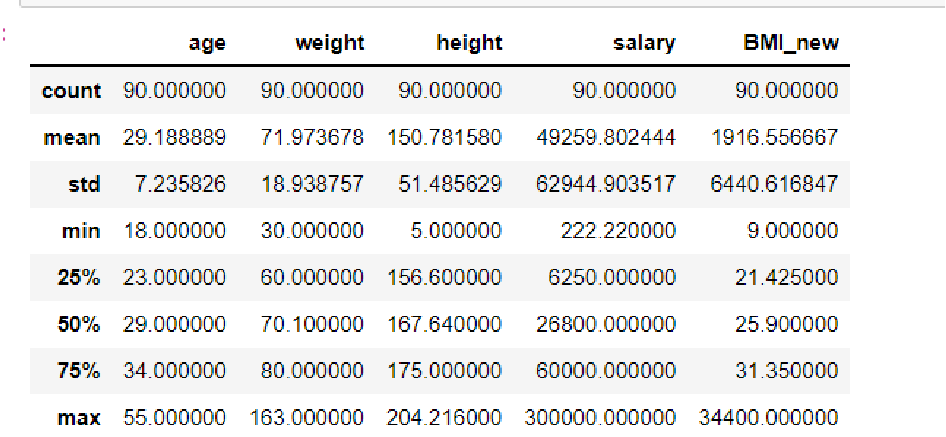
| | age | city | weight | height | BMI | salary | gender | qualification | employment_status | marital_status | favourite_dish |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 31.0 | Karachi | 73.0 | 165.0 | 26.8 | 10000.0 | Male | Bachelors | Self Emplyed | Married | Biryani |
| **1** | 21.0 | Turkey | 70.0 | 170.0 | 24.9 | 3000.0 | Male | Intermediate | Unemployed | Other | Biryani |
| **2** | 41.0 | Faisalabad | 72.5 | 156.0 | 31.2 | 35000.0 | Male | Bachelors | Employed | Single | Daal |
| **3** | 22.0 | Rawalpindi | 75.0 | 155.0 | 32.0 | 50000.0 | Male | Bachelors | Self Emplyed | Single | Karahi Ghosht |
| **4** | 26.0 | Harbin | 75.0 | 179.0 | 23.4 | 35000.0 | Male | Masters | Unemployed | Single | Daal |

Using the '.info()' method, the data types for all columns was assessed to ensure they were all in a form that could be easily manipulated.



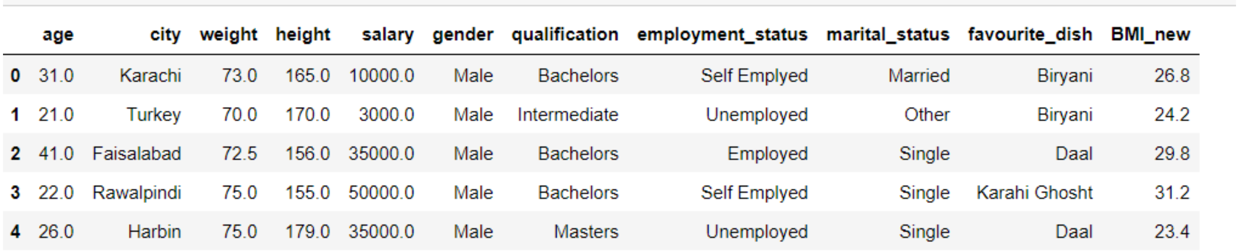
```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 90 entries, 0 to 89
Data columns (total 11 columns):
 #   Column             Non-Null Count  Dtype  
---  ------             --------------  -----  
 0   age                90 non-null     float64
 1   city               90 non-null     object 
 2   weight             90 non-null     float64
 3   height             90 non-null     float64
 4   salary             90 non-null     float64
 5   gender             90 non-null     object 
 6   qualification      90 non-null     object 
 7   employment_status  90 non-null     object 
 8   marital_status     90 non-null     object 
 9   favourite_dish     90 non-null     object 
 10  BMI_new            90 non-null     float64
dtypes: float64(5), object(6)
memory usage: 7.9+ KB
```

Columns with incorrect data type were changed. Age column was changed to int 64 data type.
Using the '.describe()' method, a summary of the data was given and it was easy to tell if values were making sense or not.



| | age | weight | height | salary | BMI_new |
|---|---|---|---|---|---|
| **count** | 90.000000 | 90.000000 | 90.000000 | 90.000000 | 90.000000 |
| **mean** | 29.188889 | 71.973678 | 150.781580 | 49259.802444 | 1916.556667 |
| **std** | 7.235826 | 18.938757 | 51.485629 | 62944.903517 | 6440.616847 |
| **min** | 18.000000 | 30.000000 | 5.000000 | 222.220000 | 9.000000 |
| **25%** | 23.000000 | 60.000000 | 156.600000 | 6250.000000 | 21.425000 |
| **50%** | 29.000000 | 70.100000 | 167.640000 | 26800.000000 | 25.900000 |
| **75%** | 34.000000 | 80.000000 | 175.000000 | 60000.000000 | 31.350000 |
| **max** | 55.000000 | 163.000000 | 204.216000 | 300000.000000 | 34400.000000 |

The min height and min and max BMI were not making sense. So, the height values below 100cm were replaced with the average height (150.78cm) in the data set.
Looking at the data, I noticed that the BMI column was not accurate, so a new column was created using the BMI equation this is why the old BMI column was dropped.

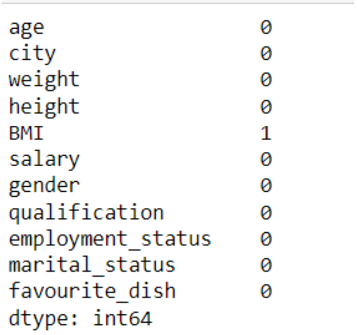


| | age | city | weight | height | salary | gender | qualification | employment_status | marital_status | favourite_dish | BMI_new |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **0** | 31.0 | Karachi | 73.0 | 165.0 | 10000.0 | Male | Bachelors | Self Emplyed | Married | Biryani | 26.8 |
| **1** | 21.0 | Turkey | 70.0 | 170.0 | 3000.0 | Male | Intermediate | Unemployed | Other | Biryani | 24.2 |
| **2** | 41.0 | Faisalabad | 72.5 | 156.0 | 35000.0 | Male | Bachelors | Employed | Single | Daal | 29.8 |
| **3** | 22.0 | Rawalpindi | 75.0 | 155.0 | 50000.0 | Male | Bachelors | Self Emplyed | Single | Karahi Ghosht | 31.2 |
| **4** | 26.0 | Harbin | 75.0 | 179.0 | 35000.0 | Male | Masters | Unemployed | Single | Daal | 23.4 |

Checks were done for duplicate rows. No duplicate rows were found.

## MISSING DATA

Once columns that were not relevant for the analysis were removed. In order to determine how many missing data each column had, Pandas '.isna().sum()' method was used.

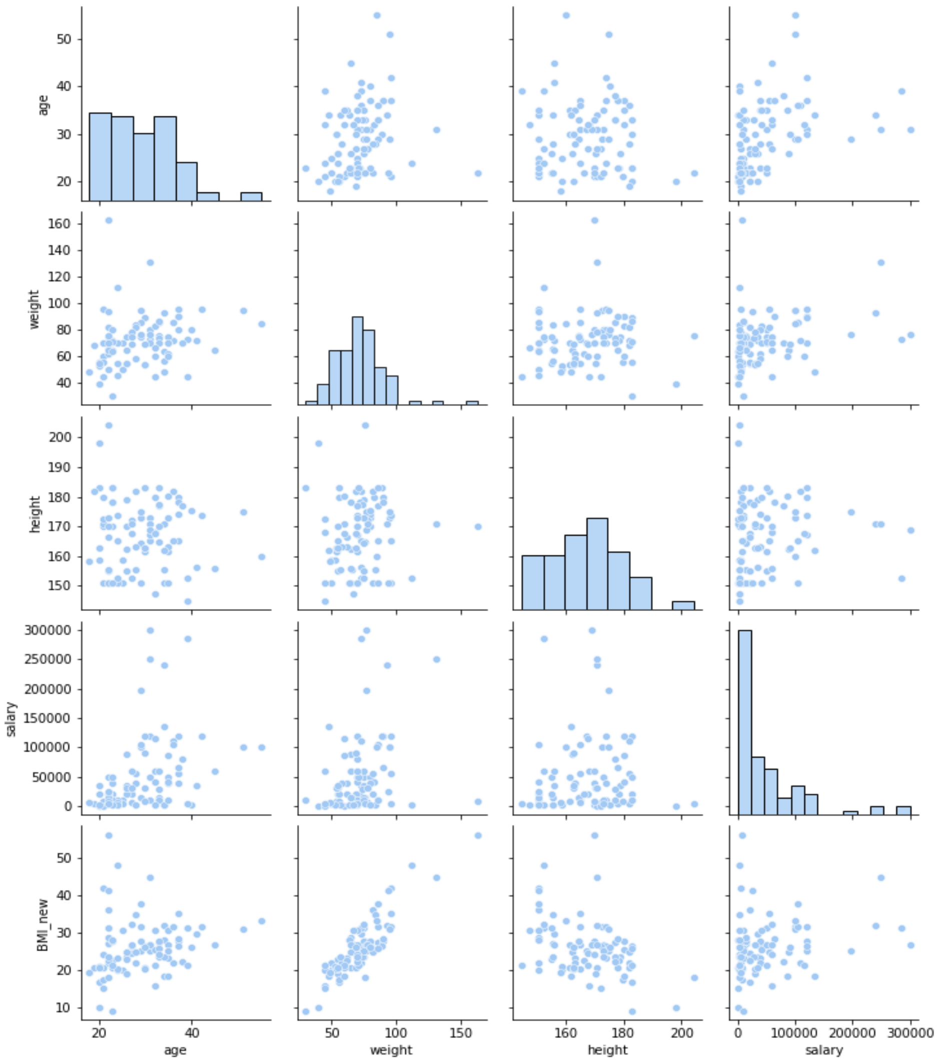


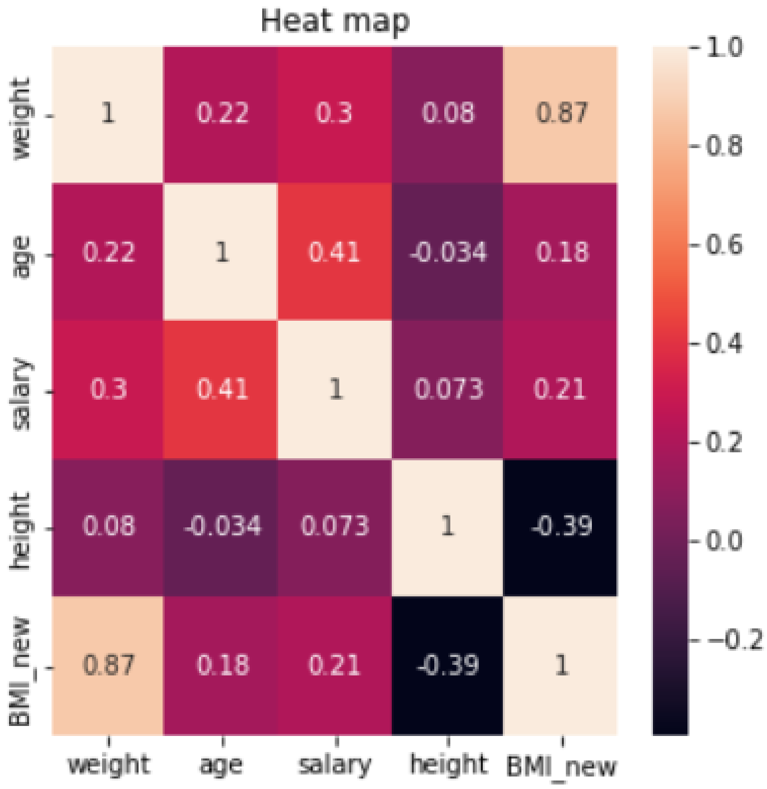
```
age                  0
city                 0
weight               0
height               0
BMI                  1
salary               0
gender               0
qualification        0
employment_status    0
marital_status       0
favourite_dish       0
dtype: int64
```

There was one missing data in the old BMI column. However, that was addressed by recalculating the BMI using weight and height.

## DATA STORIES AND VISUALISATIONS

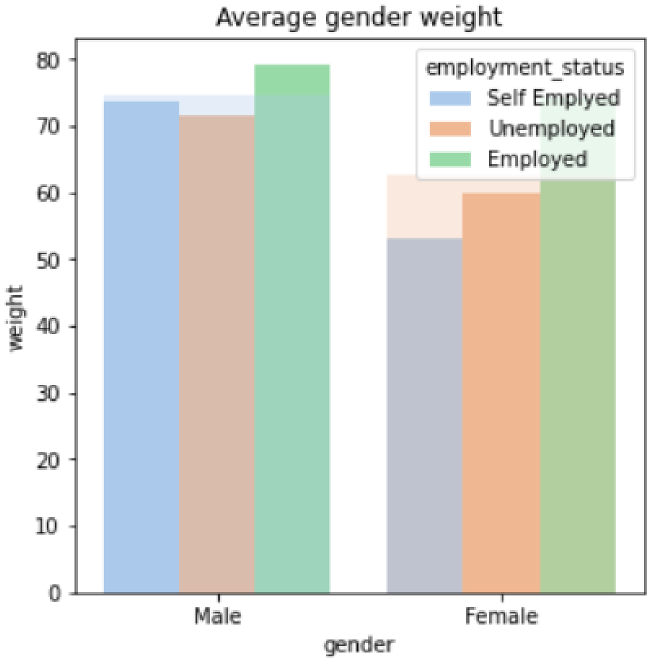
The data was visualized in a pair plot to get a general idea of related attributes.



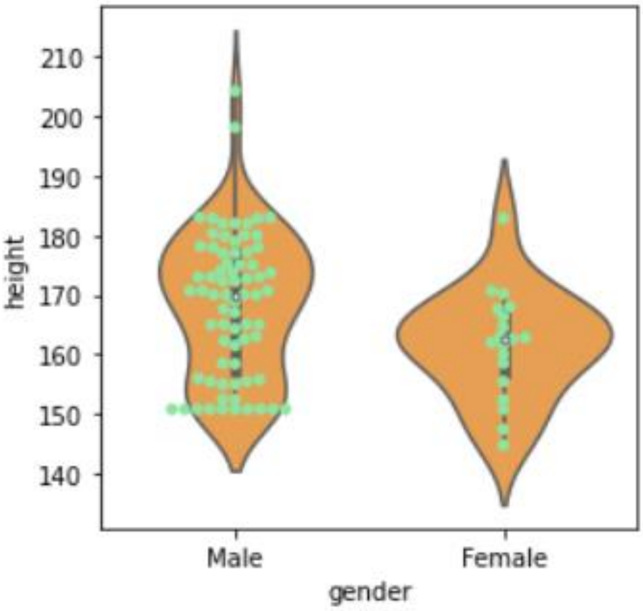


From the pair plot and heat map, there is a strong positive relationship with weight and BMI_new. This means, the more you weigh, the higher your BMI, which puts you at risk of being overweight or obese if you are not careful. Salary and age have a moderately positive relationship. The older you get, the more experience and qualifications you have, which provides a higher chance of increasing your earning potential.

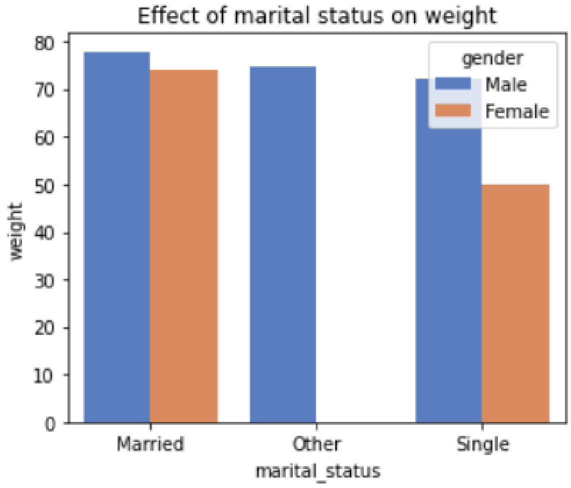
Further investigation into the dataset showed that on average male participants weighed more that the female participants. Also, having a constant source of income (i.e. being an employee as opposed to being an entrepreneur) resulted in more weight across both genders. This may be as a result of less activity/mobility with desk jobs and more money to spend on food or eating out.



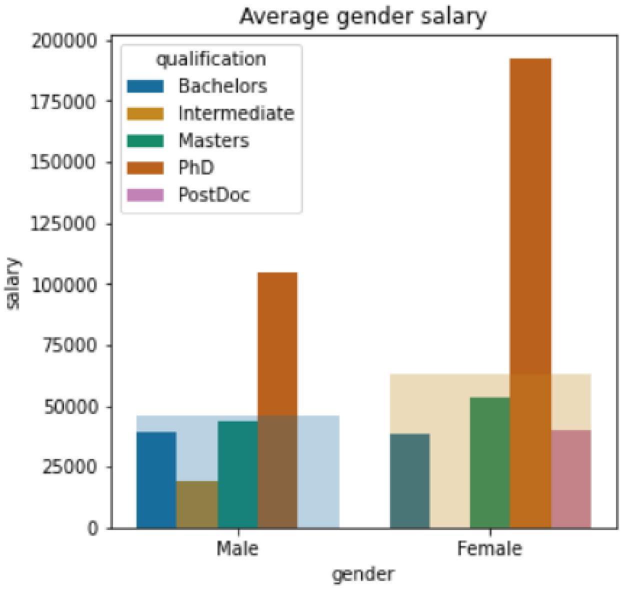
Men are usually taller than women, as shown be the data. Men also have more muscle than women, and because muscle burns more calories than fat, men tend to have faster metabolism (washingtonpost.com). Increased metabolism also increases appetite and muscle weighs more than fat, so it makes sense that men also weigh more than women.



The data also shows that married people weigh more than single people. This could be happy weight, as people usually gain weight in relationships. People get comfortable and put less focus on their appearance, while single people keep up their appearance to attract a mate. Another reason for the weight gain in married people could be an increase in salary between the two parties compared to one income for single people.

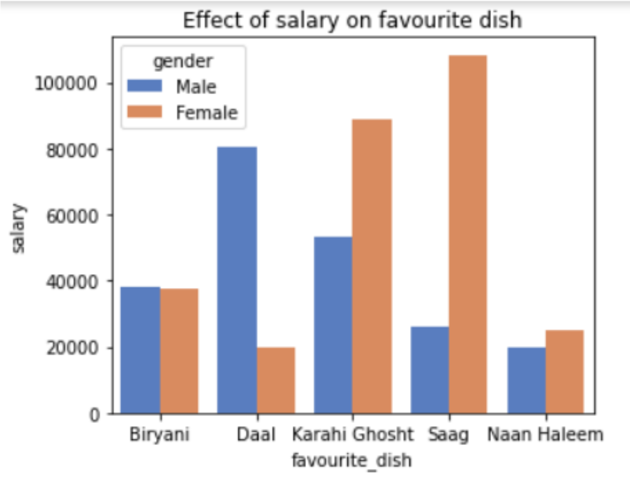


The data showed that women on average earn more than men. This is because women tend to pursue higher education more than men and based on this data, salary is proportional to qualification. However, it appears that going beyond a PhD does not have much financial benefit.

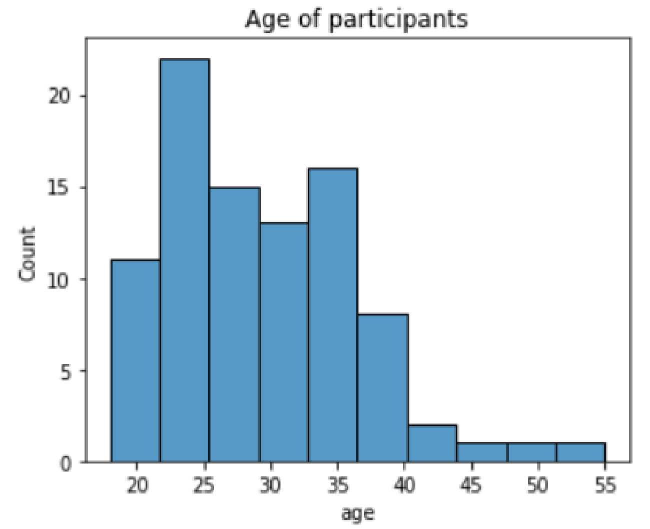


The more money earned seems to affect favourite dish. This probably means the price of these dishes are proportional to salary or that there may be some stereotypes associated with certain dishes.
Higher earning men prefer Daal (Lentils/beans dish) and Karahi Ghosht (Lamb curry) and higher earning women prefer Saag (Vegetable dish) and Karahi Ghosht.
Women probably stay away from Daal, as it is a bean/lentils-based meal which is high in fibre and causes gas.

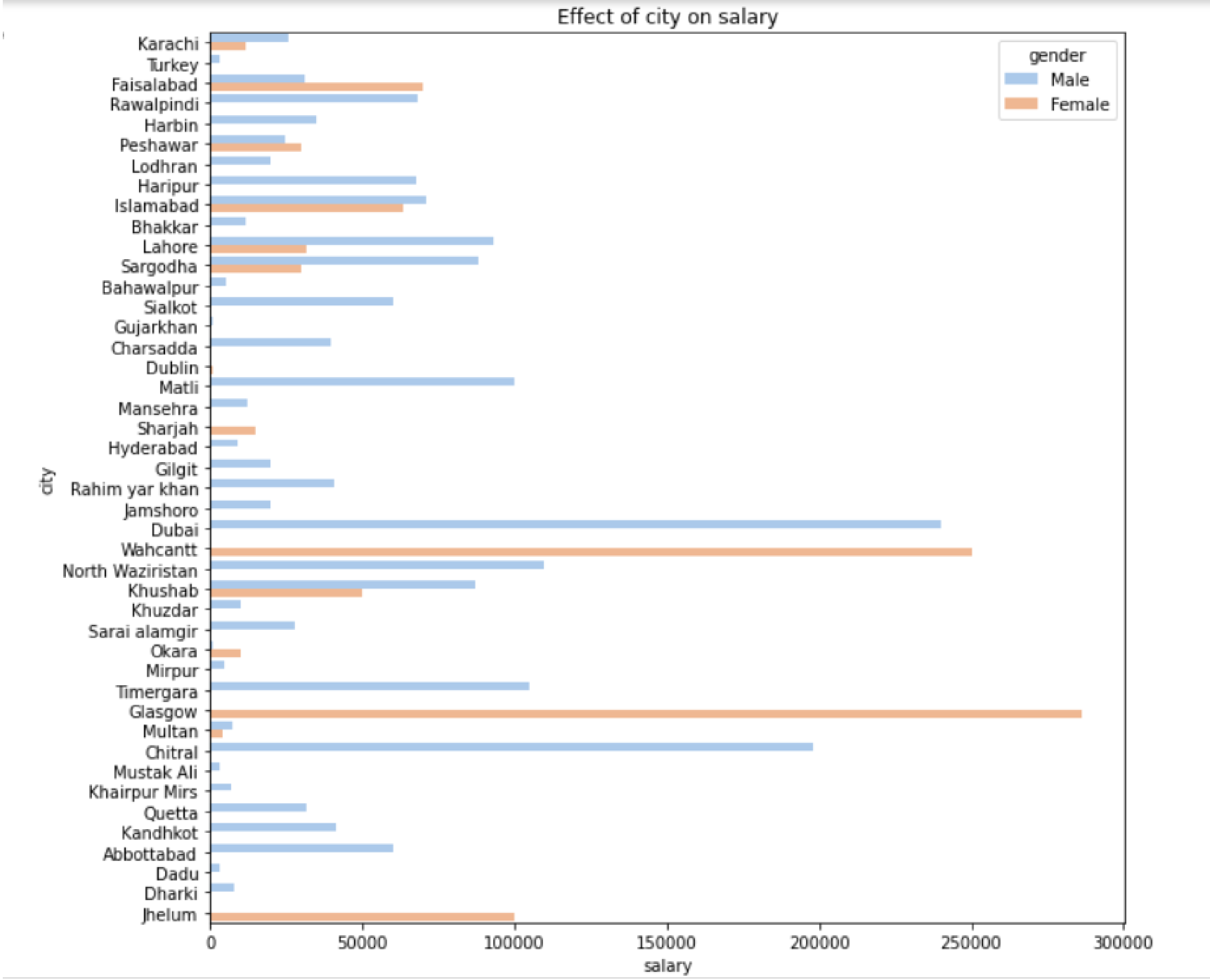
Biryani (Rice/meat/potatoes) and Naan Haleem (lentil and roti/bread) are popular in low-income earners. This shows that these meals are affordable and easily accessible.



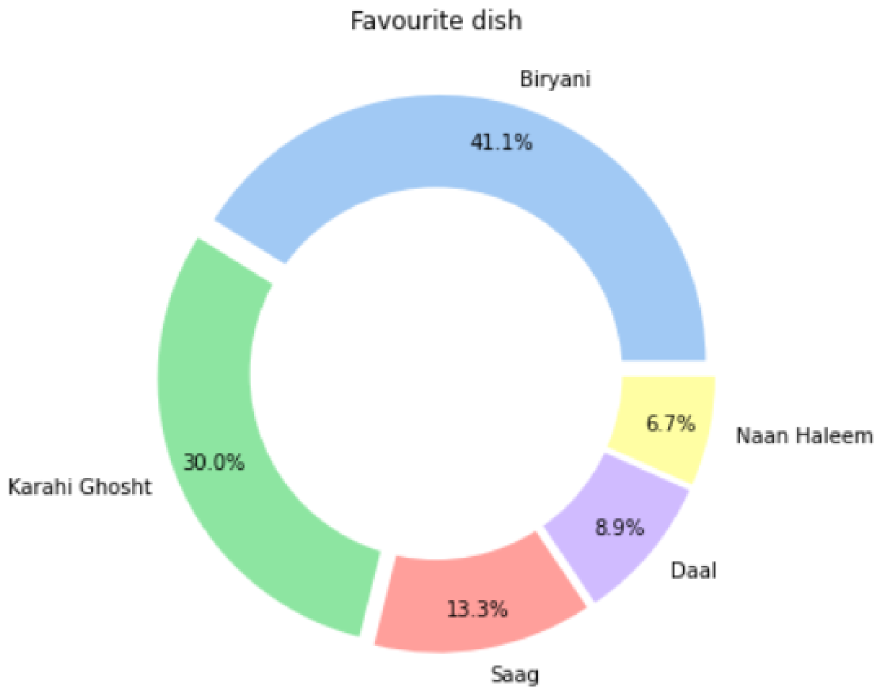
The participants ages ranged between 18-55. Weight is also affected by age, as metabolism slows down as you age, so more attention should be paid to diet and exercise as you get older.



The city you live in will also affect your lifestyle. Some cities have higher costs of living than others.  Expensive cities usually have higher salaries to offset the costs. Dubai, Wahcantt and Glasgow have the highest earners (above 200000) amongst the participants and are younger than 40 years of age.



Faisalabad, Lahore and Jhelum have participants older than 40 and they earn less than 100000.



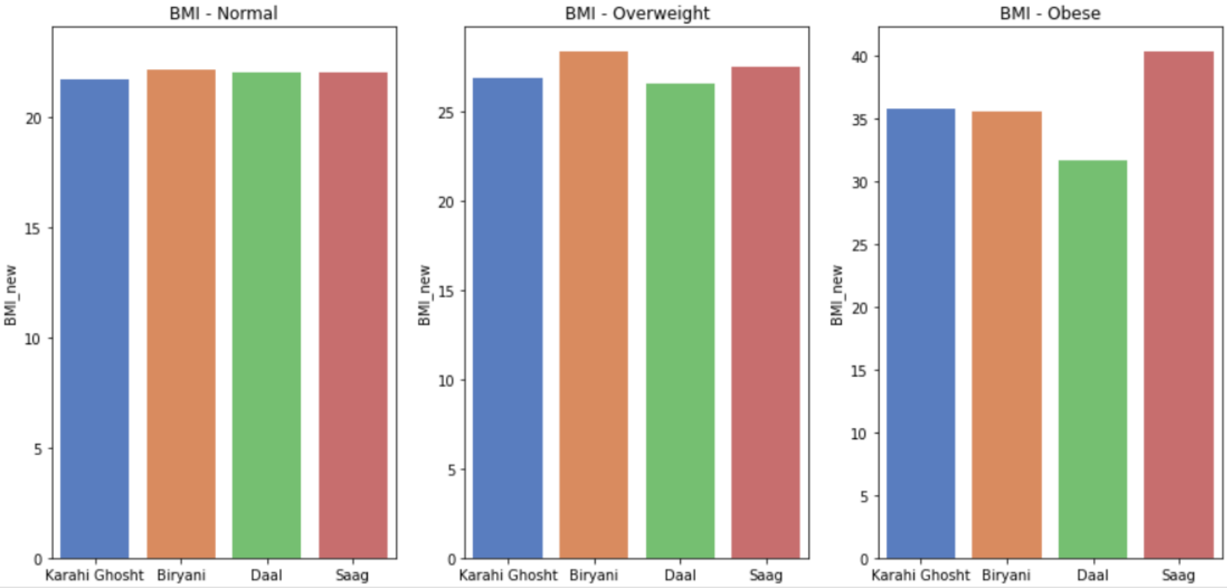
The most popular dish amongst participants was Biryani. This is a common dish and is preferred by those with lower income as it is affordable. Most cities have low-income earners.

Dishes popular with low-income earners are usually high in carbohydrates and require more effort to burn calories. As mentioned previously, the metabolism slows down over time.  However, lower income earners are more active (limited modes of transport) and may likely eat less frequently due to lack of resources.

Saag was more popular in the high-income earners and older group. Saag is an Indian leaf vegetable dish eaten with bread such as roti or naan, or in some regions with rice (Wikipedia, 2022).

As much as it is a vegetable dish, it is eaten with bread or rice which are high in carbohydrates and older people already have slower metabolism and the higher earners are not as active as they may have desk jobs or different modes of transportation and delivery services which puts them at a higher risk of gaining weight.  Women are also known to have slower metabolisms than men as stated previously.

Biryani being popular with participants under 30 years old also helps to lowers the risk of weight gain but if little to no activities are carried out weight gain is inevitable.



**THIS REPORT WAS WRITTEN BY: KP USEH**

## REFERENCES

https://www.kaggle.com/datasets/abdulraheem625/food-choice-male-vs-female-indianpakistani-food

https://www.natureschoice.co.za/bmi-calculator/

https://www.geeksforgeeks.org/donut-chart-using-matplotlib-in-python/

https://www.washingtonpost.com/lifestyle/wellness/weight-loss-it-really-is-harder-for-women-research-shows/2014/08/12/0a95c1aa-1d9b-11e4-ab7b-696c295ddfd1_story.html

https://en.wikipedia.org/wiki/Saag