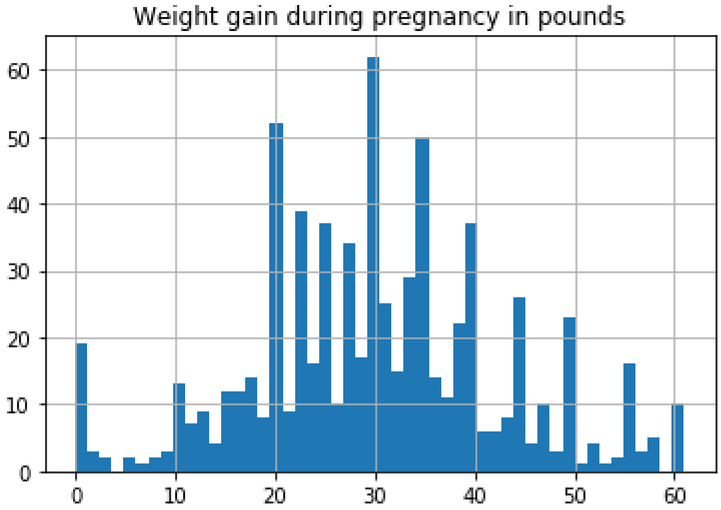
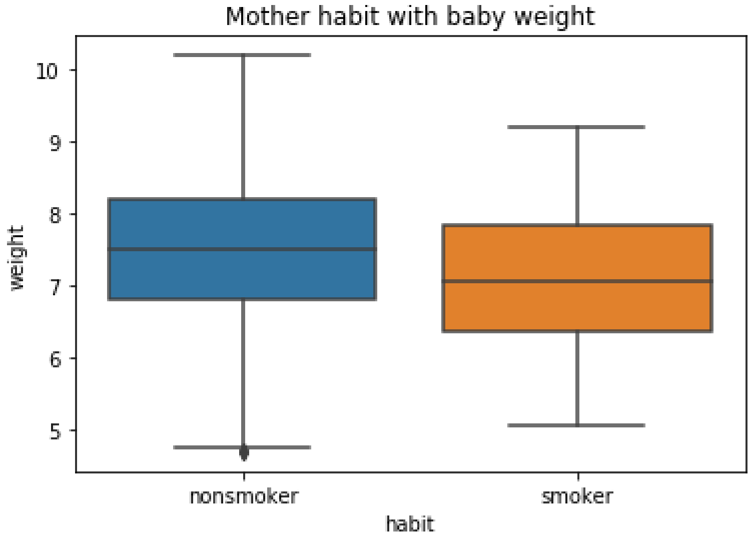
Author: Abel Stanley
NIM: 13517068

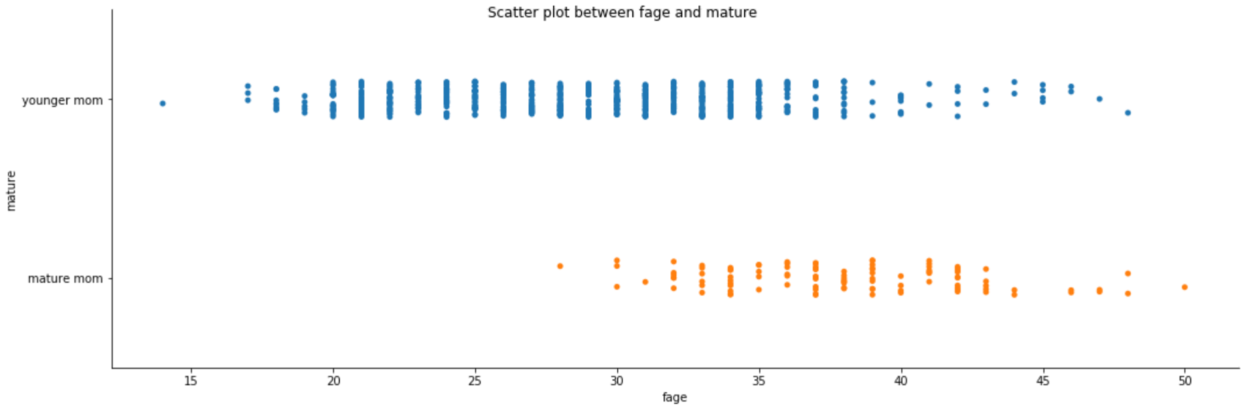
1. Interesting things about the data:
    a. Data distributions
        i. Young mom number out values mature mom number by a huge margin
        ii. There is a considerable number of mothers reporting no weight gain during pregnancy, which I think is bizarre? Impossible, no?



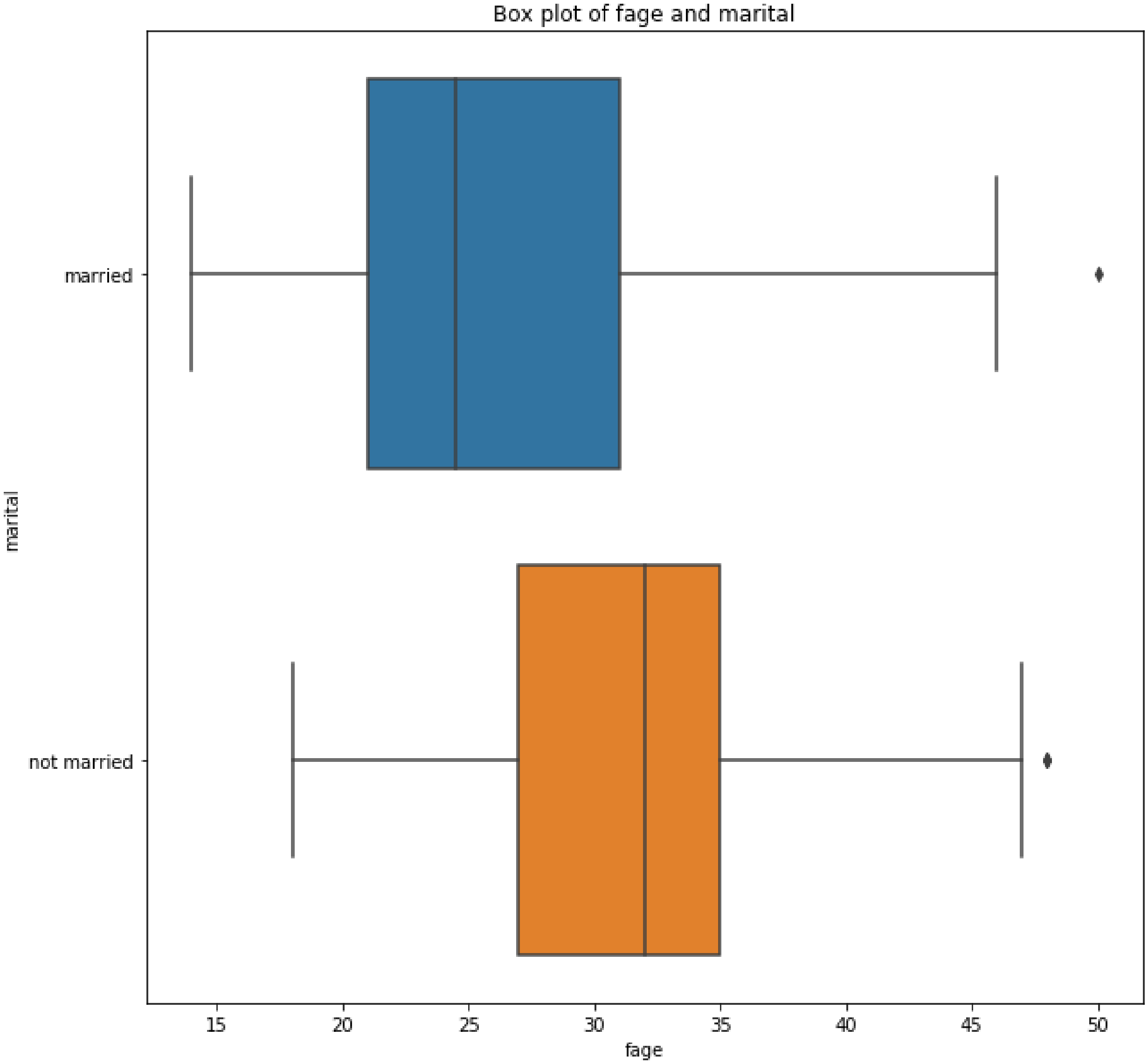
    b. Data relationship
        i. Mother who smokes tends to give birth to babies with less weight



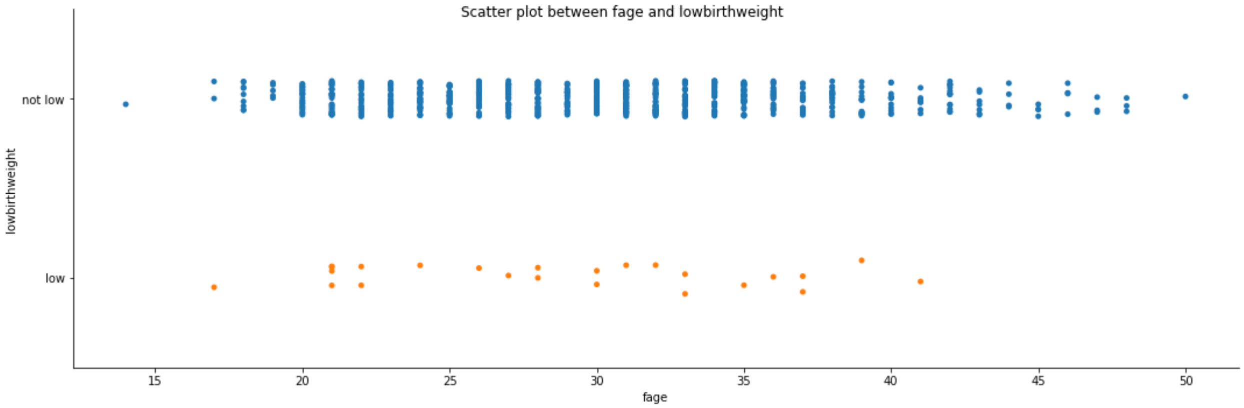
        ii. Mature moms don’t have children with fathers below 30 years old



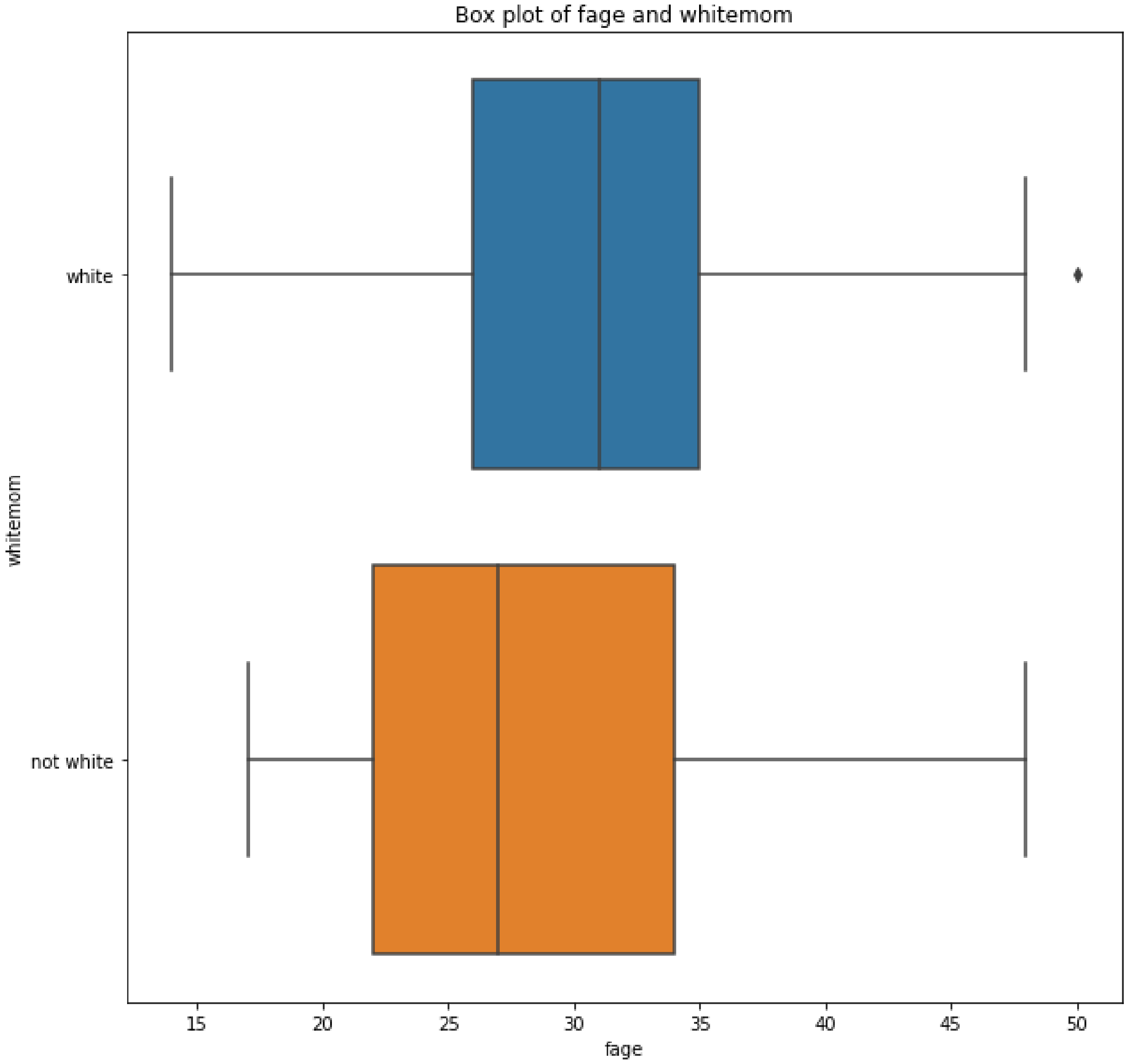
iii. Younger fathers tend to be married when having a child. Older fathers (> 30 years old) tend to not be married when having a child.



iv. Fathers at the age above 42~ has 0 occurrence of having a low weight (premature) baby



v. White mom prefers older men while non-white mom prefers younger men as husband



c. Correlation:

1. Father’s age and Mother’s age is strongly positively correlated with each other (0.77)
2. Premmie & Weeks are correlated with each other (0.56)
3. Weight & Low Birthweight are correlated with each other (0.4)
4. Weeks & Low Birthweight are correlated with each other (0.2)
5. Mother age and Mature are correlated with each other (0.64)
6. Weeks & Weight are correlated with each other positively (0.31)
7. Mother age and Marital are correlated with each other positively (0.4)

      8. Father age and Marital are correlated with each other (0.34)
      9. Father age and Mature are correlated with each other positively (0.49)

   d. Q-Q Plots
      i. From all numerical columns, the ones that have normal distribution are:
         1. Weeks
         2. Weight

   e. Distance Matrix
      i. From the distance matrix, we can see 401 and 351 is separated by a considerable amount of distance from the rest of the distribution
      ii. Estimation: 51,151, 201, 251,301, 451, 501, 551, 601, 651 are closely located with each other
      iii. 451, 501, and 151 are possible centroids for a cluster
      iv. 1 and 701 are located moderately far from the rest of the distribution

   f. Other interesting facts:
      i. White mother who smokes (0.10460992907801418%) is more common than non-white mother who smokes (0.07792207792207792).
      ii. White mother prefers to have female babies (0.5177304964539007%) and Non-white mother prefers to have male babies (0.5454545454545454%)
      iii. White mothers tend to have children before they are married (0.22340425531914893%) while non-white mothers are more likely to have children after they are married (0.4675324675324675%)

2. Questions for discussion
   a. Is IQR always applicable for all data distribution to detect outliers? Because even after using IQR to drop outlier rows, there are still a few outliers left visible on boxplots and heatmaps from distance matrix
   b. How to plot/visualize relationships between features if both are categorical? Box-plots, scatter plots, and Q-Q plots fail to plot data if both are categorical
   c. What is Q-plot? I can only find Q-Q plot (Quantile-quantile plots)