Yajing Zhou, Zhuo Shan

HW 3

INFO 370

**Mini Data Science Project**

In this assignment, we used athlete data from Strava and focused on analyzing the following two questions:

1. Do men tend to exercise more intensely than women?
2. For riding athletes that did exercise in their home country, does the total elevation that riding athletes gained contribute more to their heart rate than the ride distance does?

For each question, trying to get the answer, we cleaned and prepared data, did exploratory data analysis and statistical modelling. We will show that below.

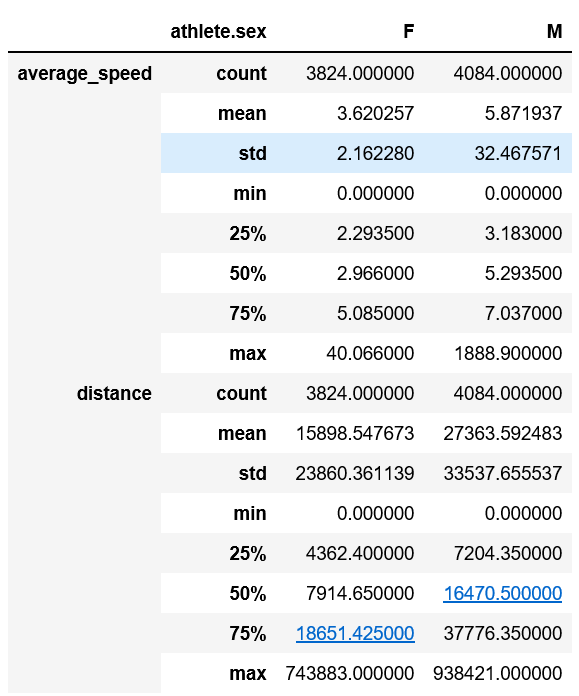
**Question 1: Do men tend to exercise more intensely than women?**

**Data Preparation**

For this question, we dragged four columns from the original dataset. They are ‘athlete.sex’, ‘average\_speed’, ‘distance’ and ‘type’. These four are interest of variables that contribute to the first question. ‘athlete.sex’ represents the gender of an athlete. ‘average\_speed’ represents the average speed of an activity in meters per second. ‘distance’ is the distance in meters and ‘type’ is the type of activity. Additionally, we removed the entire row if there is a Nan in a row.

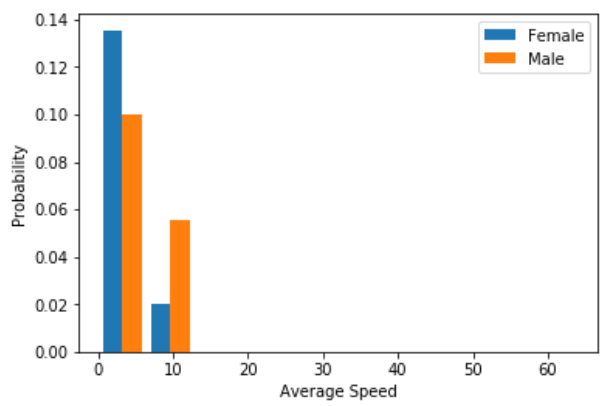
**Exploratory Data Analysis**

Firstly, we grouped the data by gender (‘athlete.sex’) and got the summary of the numeric variables, ‘average\_speed’ and ‘distance’.



It seems that the mean of male’s average speed is larger than that of female’s average speed. However, we cannot get the conclusion from this because there are other factors can lead to this situation. For example, we think that the longer the exercise distance will affect physical strength of athletes. Maybe in this data, men did more activities with short distance such that men did higher average speed.

To get rid of the effect of different distance, we dragged the data where distance ranges from 0 to 743883 that is the maximum distance of female’s activities. Then, we checked the distribution.

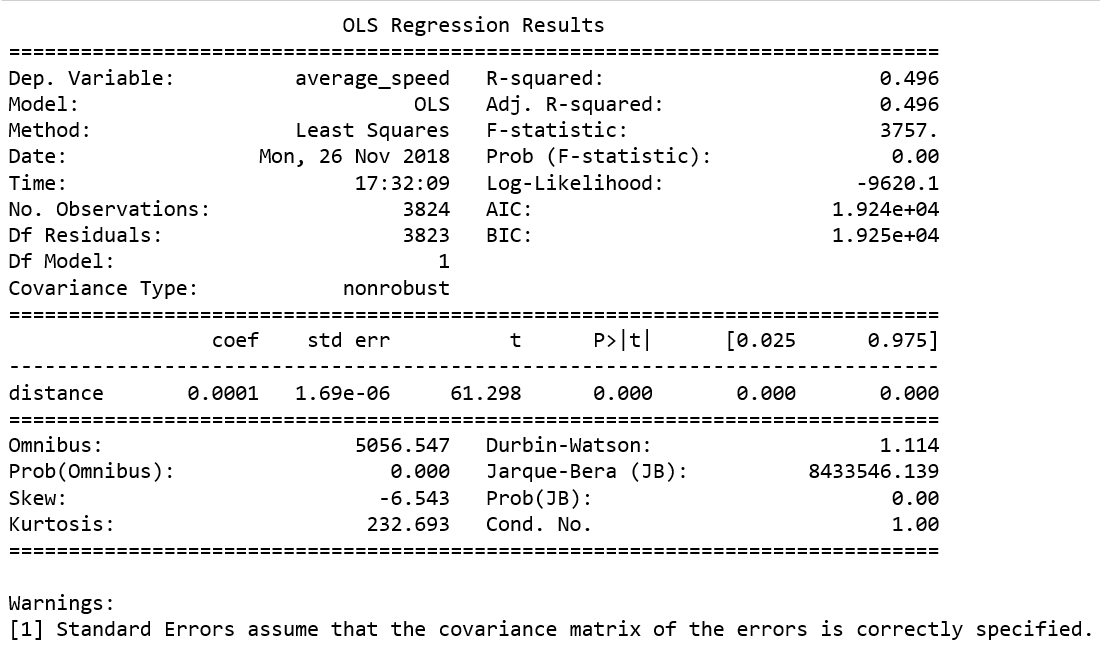


From this plot, we can see men performed higher probability in average speed. Therefore, we made a hypothesis that men tend to exercise more intensely than women.

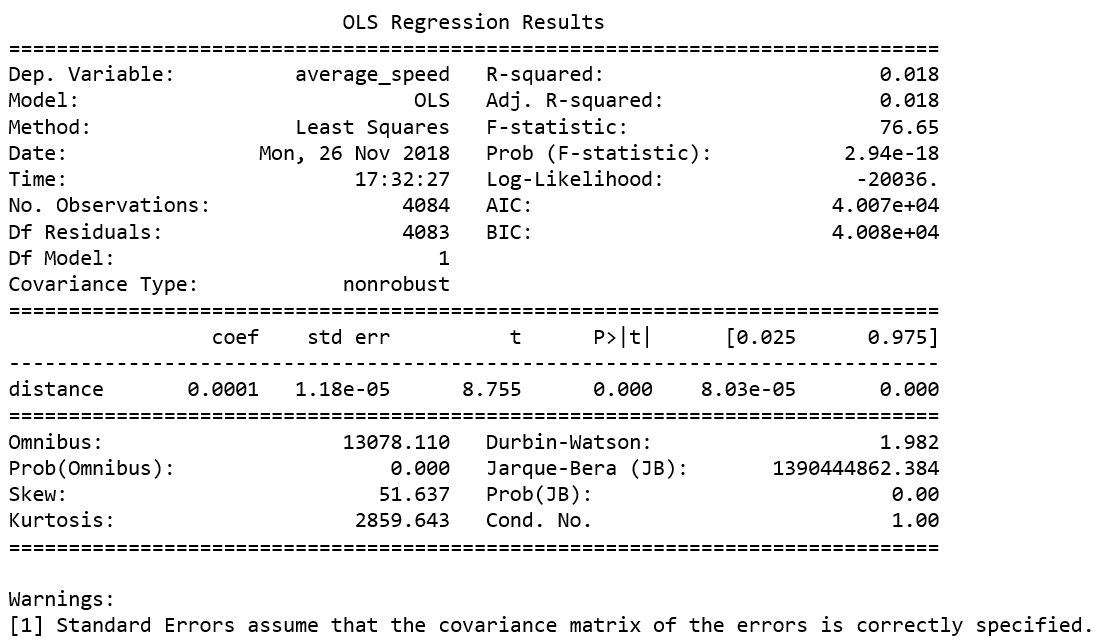
**Statistical Modelling**

To test our hypothesis, we have wanted to see the correlation between distance and average speed based on different gender. Then, we would compare the units of change in average speed once one unit of distance increase to see the intensity difference between female and male. If there is a larger increment in male’s average speed for every unit of increment in distance then that in female’s, we would get the conclusion that men tend to exercise more intensely than women. Otherwise, we would like to say that women tend to exercise more intensely than men.

Firstly, we ran the model no matter what the type of an activity is, and we got the models below.



Model for Female



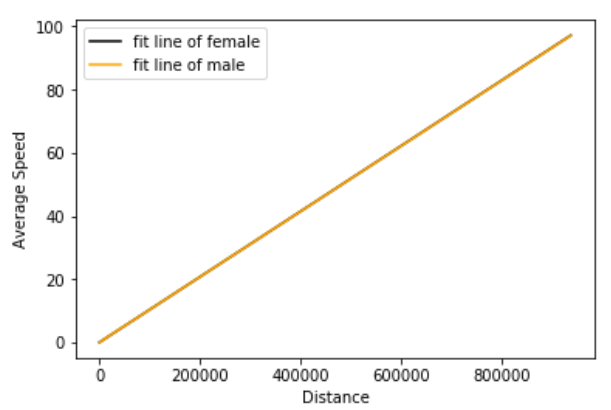
Model for Male

We did not add constraint to the independent variable here because we think that there are should not have an intercept for the model here, which means that the average speed should be 0 if the distance is 0.

**Results**

For the model for female, each meter of increase in distance will lead to 0.0001 units of increase in female’s average speed. For the model for male, each meter of increase in distance will lead to 0.0001 units of increase in female’s average speed.

The parameters for two models are closed (same) to each other. We also plotted the fitted lines of these two models.

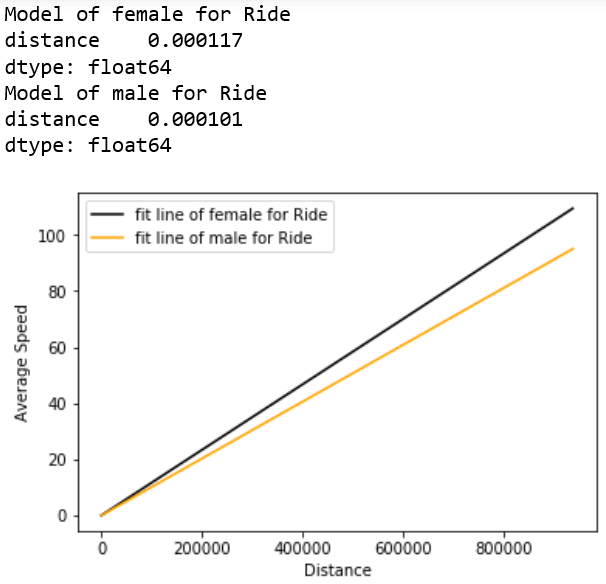
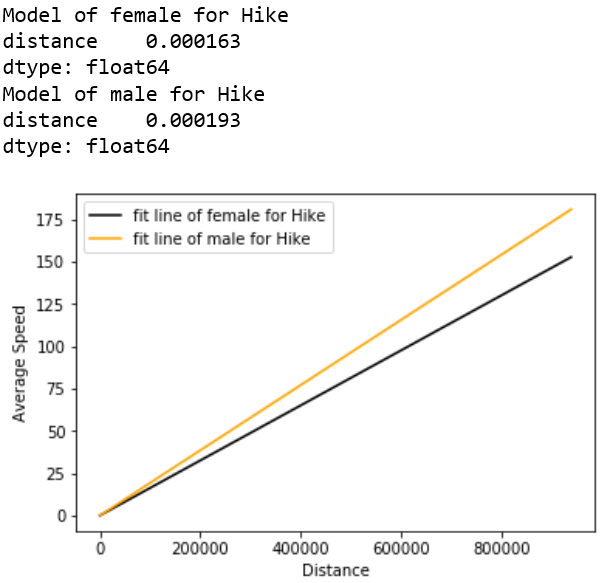
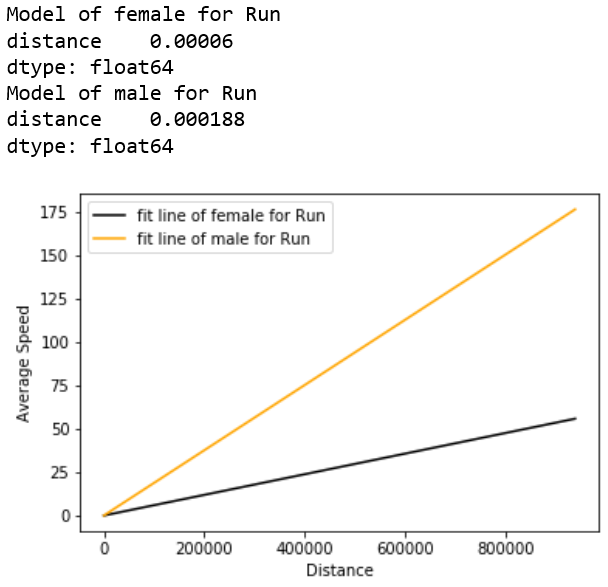
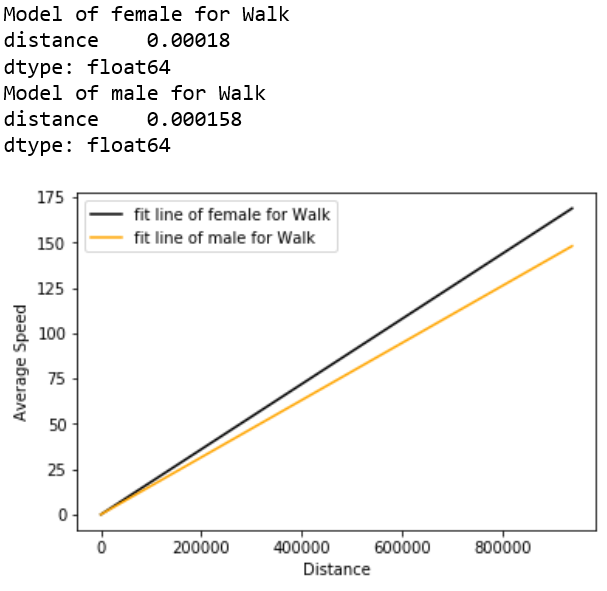


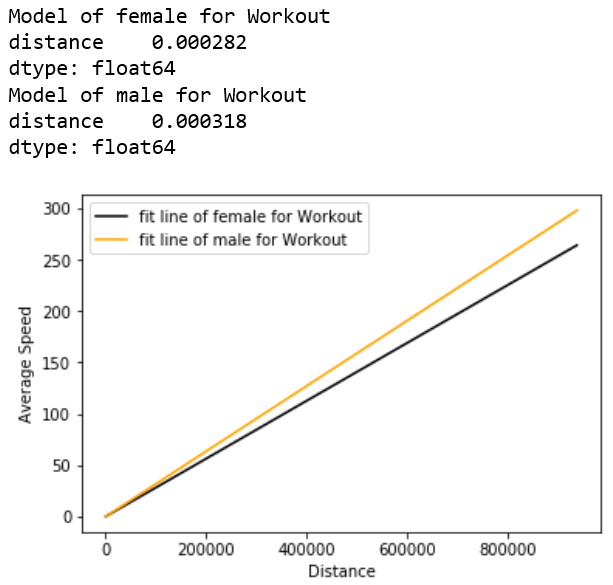
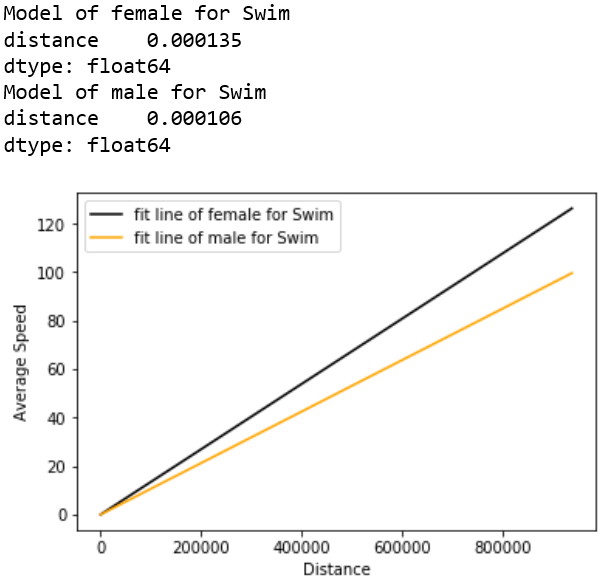
It seems that the fitted lines for them are kinds of overlapping, which means men do exercise as intensely as women.

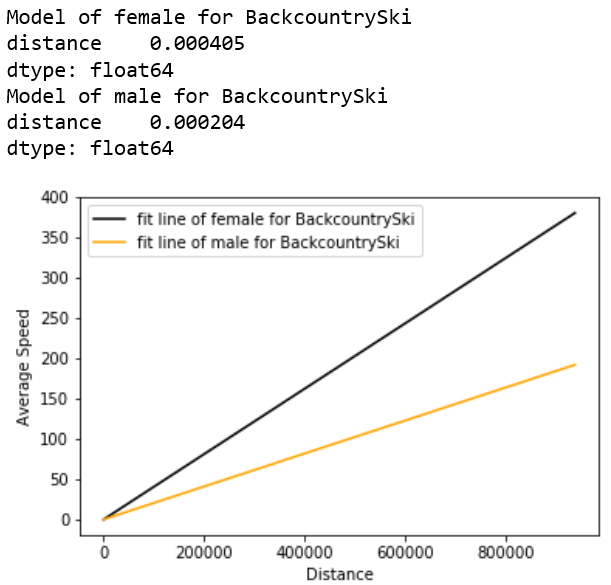
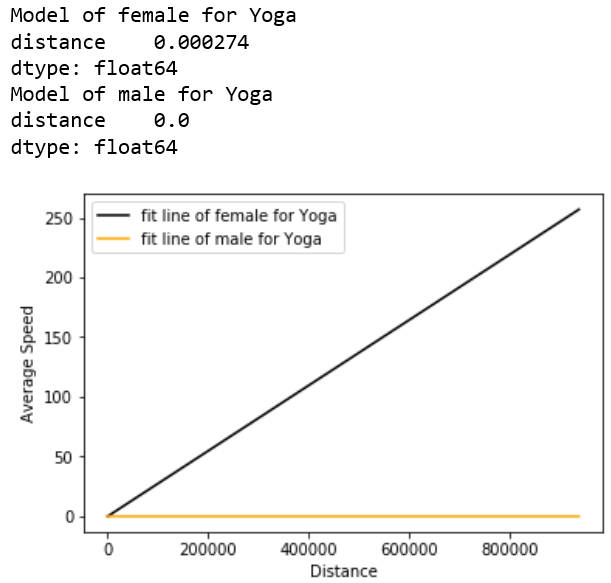
**Discussion**

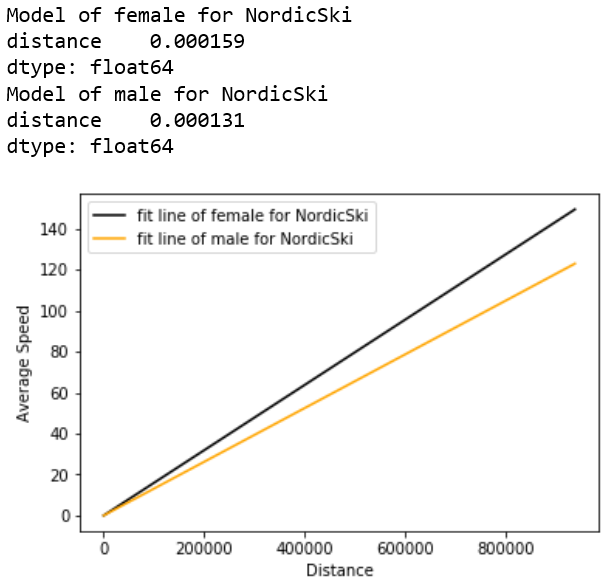
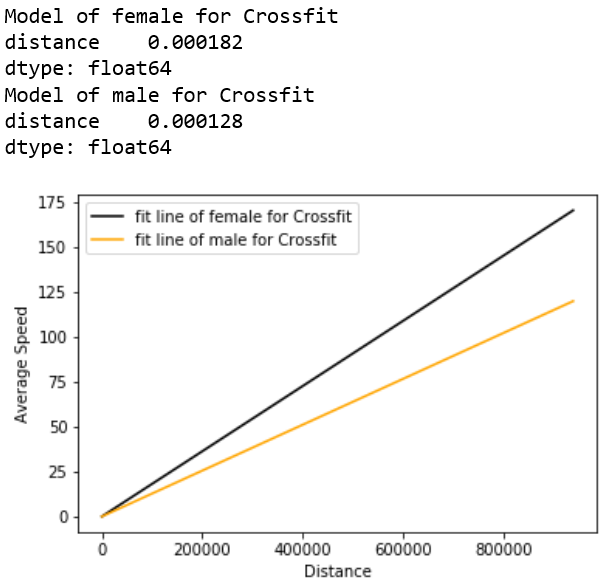
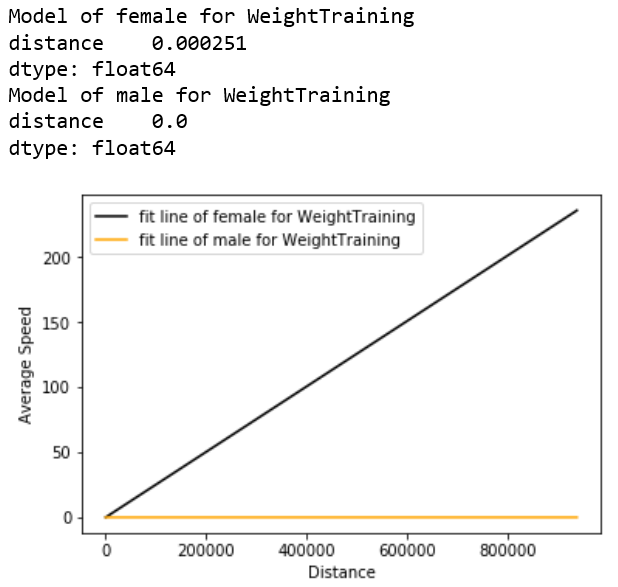
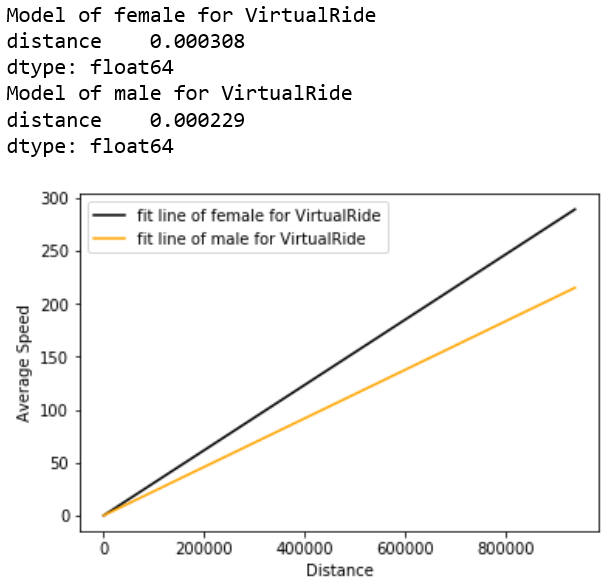
From above, we got the conclusion that there is no gender difference in exercise intensity when we are considering overall exercises. However, we are hard to say that men and women perform same level of intensity for every exercise. And, we think the reason lead to this conclusion because men do some types of exercises more intensely than women and women do other types of exercises more intensely than men and they were offset when we were considering all exercises together.

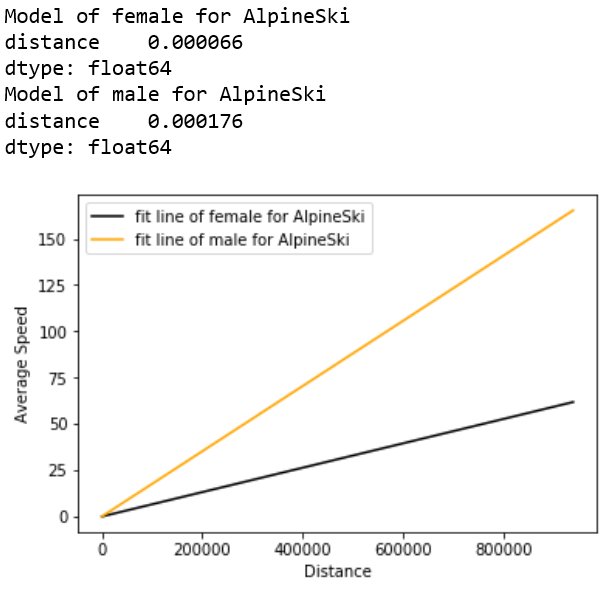
We also did same modeling process for each type of exercises.

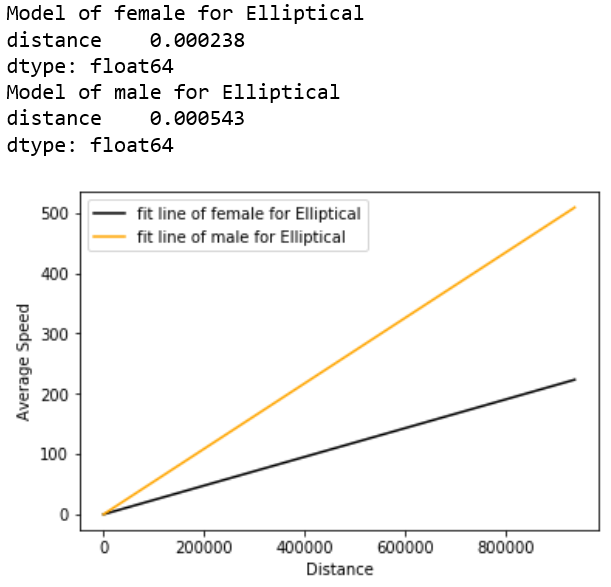
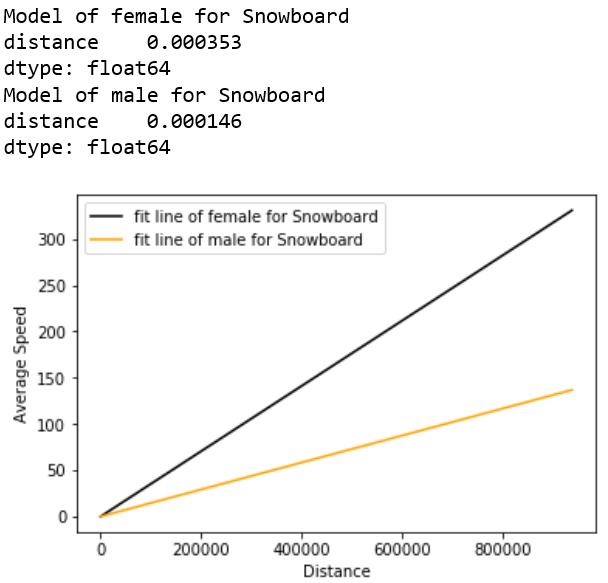
   

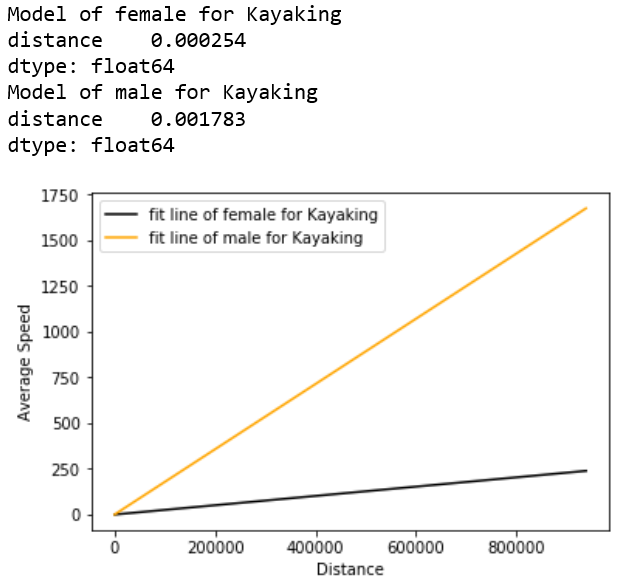
 





If the yellow line is steeper than black line in a plot of an exercise, we would like to say men do this exercise more intensely than women. If the yellow line is flatter than black line in a plot of an exercise, we would like to say women do this exercise more intensely than men.

From the plots above, for riding, walking, swimming, backcountry skiing, yoga, Nordic skiing, virtual riding, weight training, cross fitting and snowboarding, women do more intensely than men. For hiking, running, workout, alpine skiing, elliptical and kayaking, men do more intensely.

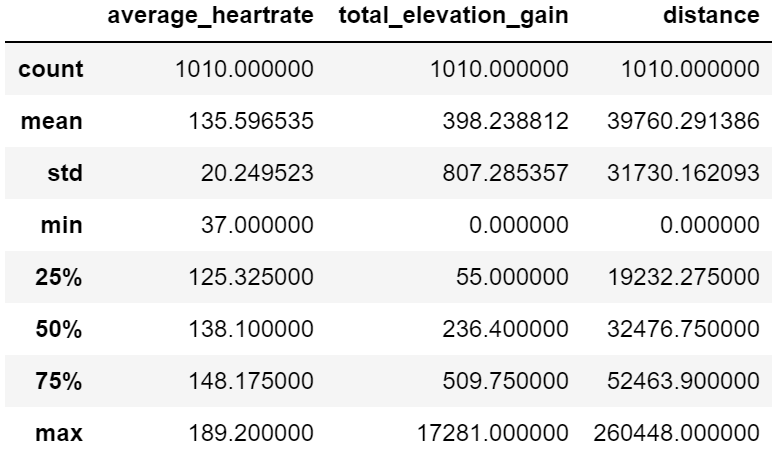
**Question 2: For riding athletes that did exercise in their home country, does the total elevation that riding athletes gained contribute more to their heart rate than the ride distance does?**

**Data Preparation**

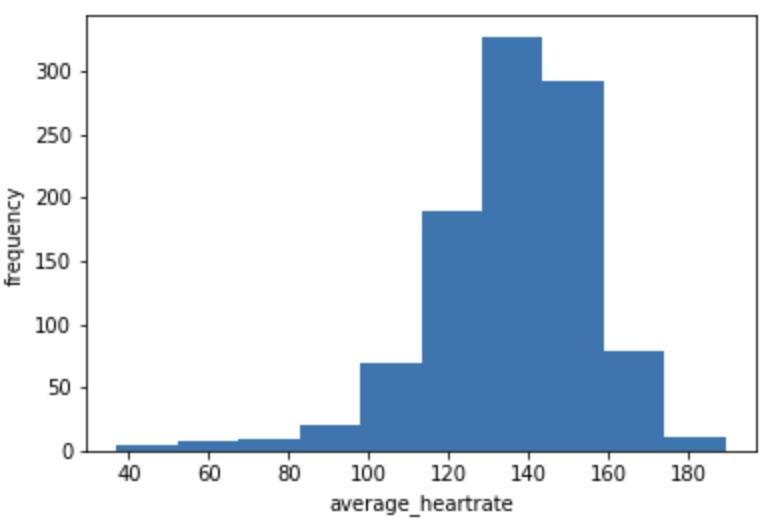
For this question, we remove observations with 0 average heart rate because it does not make sense. Then, we chose observations with the type of ride. To narrow the data down to the athletes that did exercise in their home country, we tested whether a observation has same value in the columns of ‘athlete.country’ and ‘location\_country’ and we would only use the observations with same value in the columns of ‘athlete.country’ and ‘location\_country’. After narrowing the data, we chose three columns. They are ‘average\_heartrate’, ‘total\_elevation\_gain’ and ‘distance’. ‘average\_heartrate’ is the heart rate of the athlete during this effort. ‘total\_elevation\_gain’ and ‘distance’ is used in meters. Finally, we removed Nan value from the dataset.

**Exploratory Data Analysis**

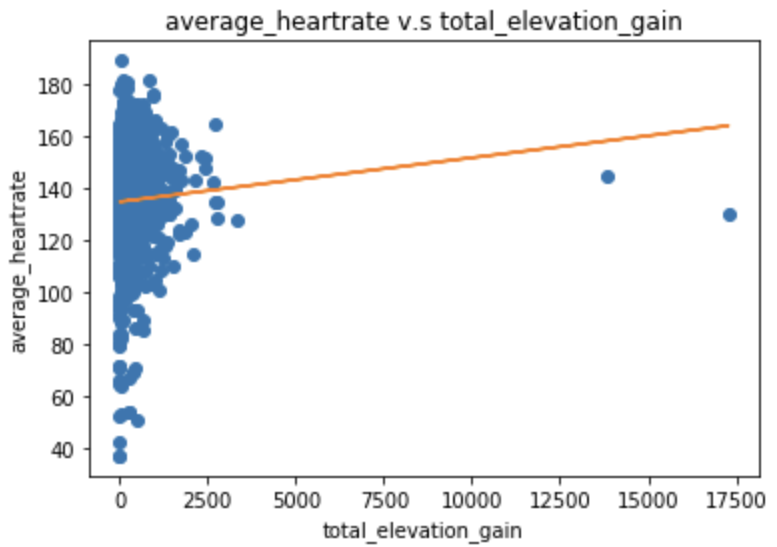
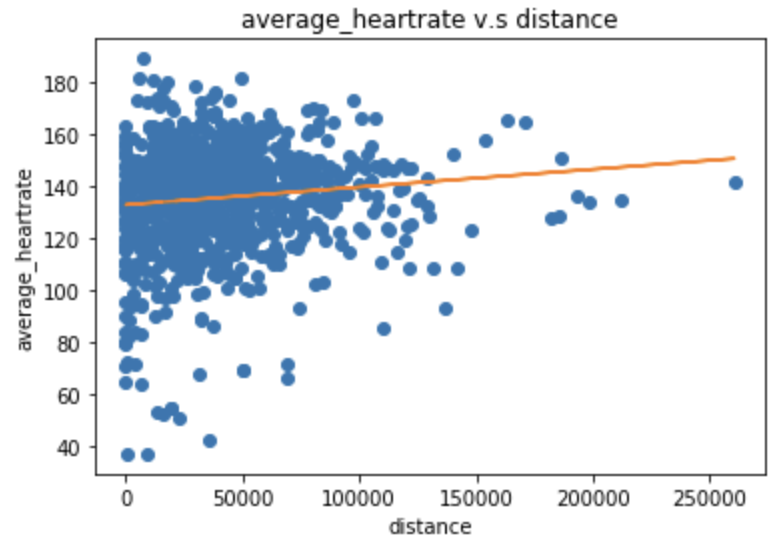
Firstly, we got the summary of this dataset.



Also, we plotted a histogram of average heart rate to see how it distributes. It seems that most of data are in the range of [100,160].



Finally, we plotted two scatter plots with fitted line. One is average heart rate v.s total elevation gain. Other one is average heart rate v.s distance.

From two plots above, we can see that both of total elevation gain and distance have positive correlation with average heartrate. It seems that the slope of average heart rate v.s total elevation gain might be larger.

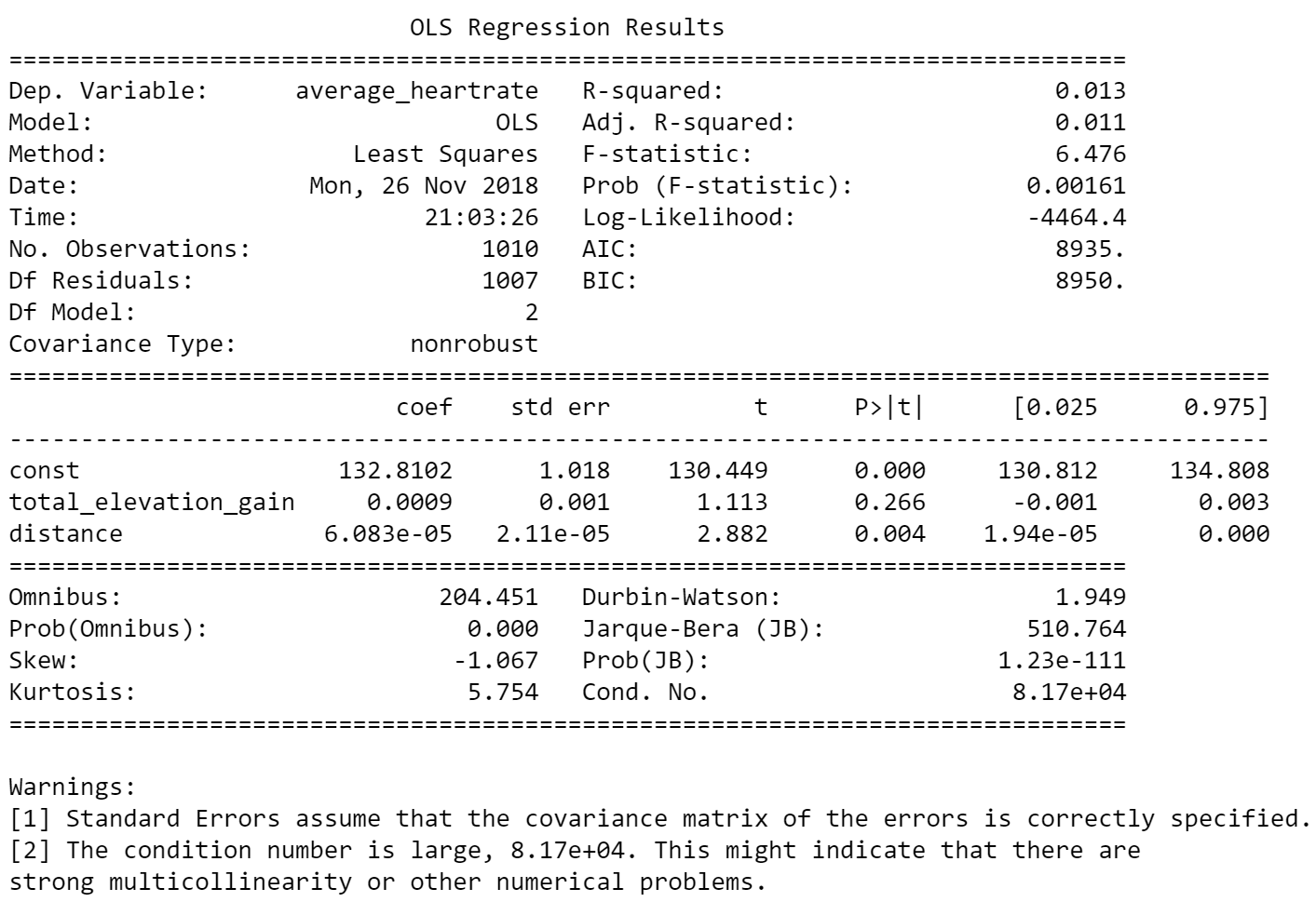
**Statistical Modelling**

From the data analysis above, we would like to set our hypothesis and our significant value is 0.05.

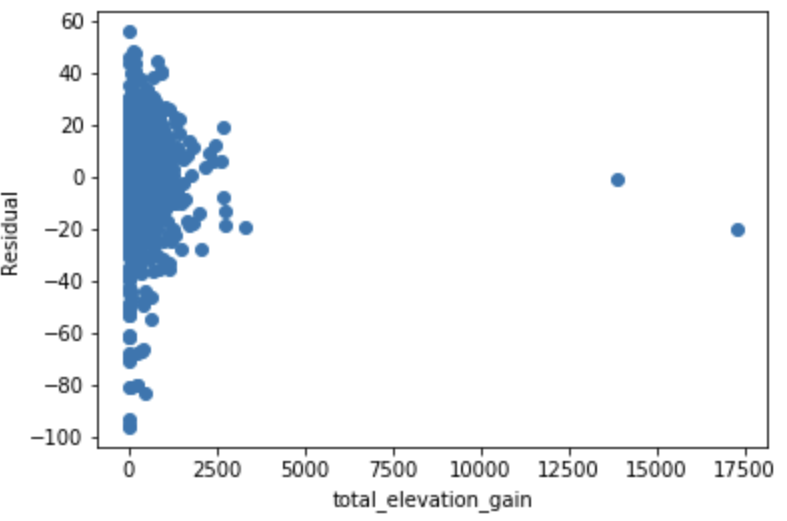
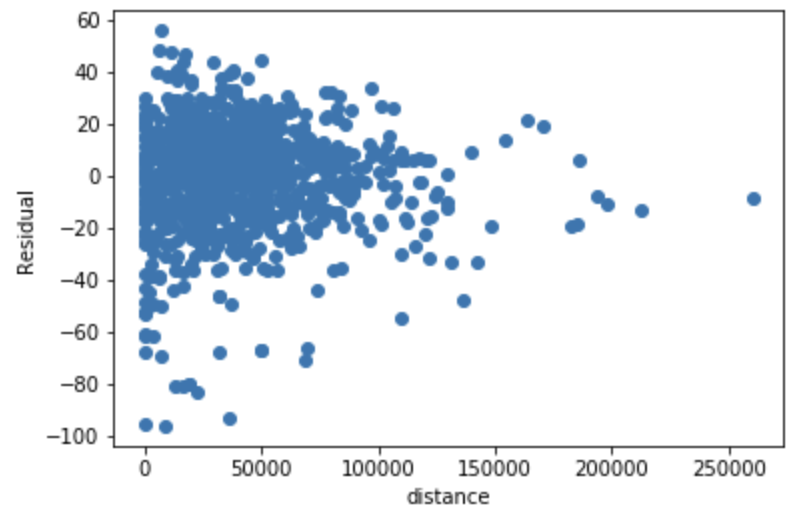
Null hypothesis: the ride distance contributes to their heart rate as same as the totalelevation gained does.

Alternative hypothesis: the total elevation ride athletes gained contributes more to their heart rate than the ride distance does.

We ran a multivariable linear regression model and we got the summary below:

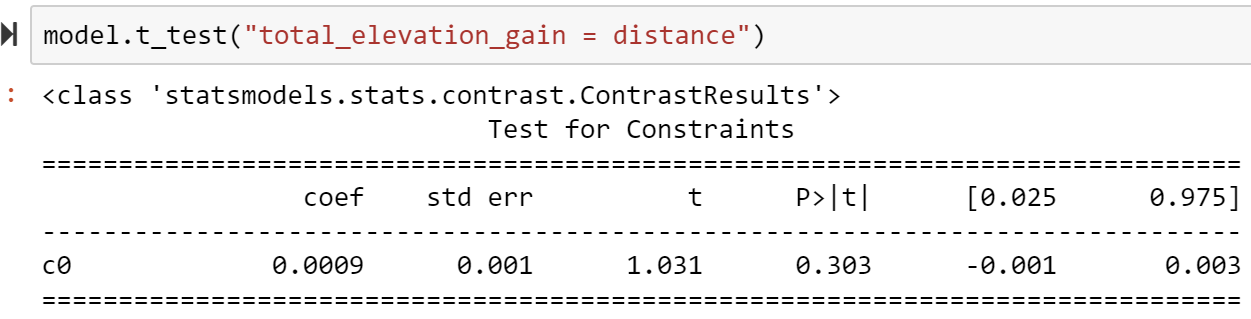


Then, we plotted scatter plots of each of independent variables and residuals.

The mean of residuals in both plots are closed to 0, which means the model reasonable.

To do hypothesis test, we ran a code below and got the feedback.



The p-value here is 0.303 which is greater than our significant value 0.05, which means we cannot reject our null hypothesis and we get a conclusion that the ride distance contributes to their heart rate as same as the totalelevationgained does.

**Results**

After checking the plots of residuals, we would like to say our model is reasonable. From the model, we know for each unit of increase in total elevation gain would lead to 0.0009 units of increase in average heartrate, remaining other variables unchanged, and for each unit of increase in distance would lead to 6.083e-05 units of increase in average heartrate, remaining other variables unchanged. And, when both of total elevation gain and distance are 0, the average heartrate is 132.8102.

After doing hypothesis test, we get a conclusion that the ride distance contributes to their heart rate as same as the totalelevationgained does.

**Discussion**

To our surprise, the effect of ride distance and total elevation gained to average heart rate is same. We used data analysis to set out hypothesis and ran statistical model to test the hypothesis. Maybe for future improvement, we can add more variables to the model to make the answer more fair.