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.

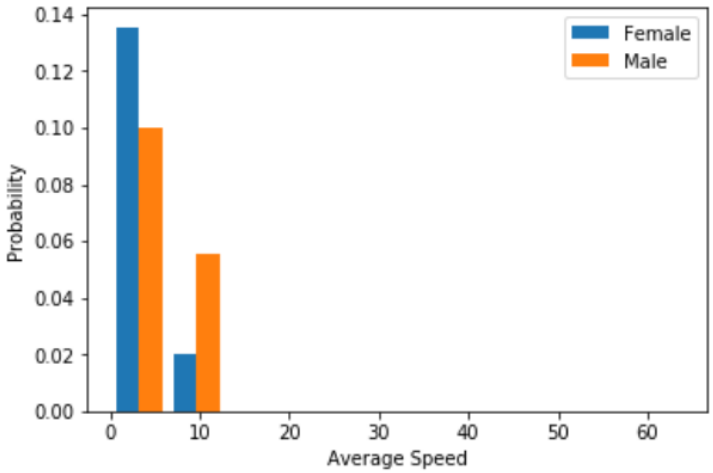
Exploratory Data Analysis

Firstly, we grouped the data by gender ('athlete.sex') and got the summary of the numeric variables, 'average_speed' and 'distance'.

	athlete.sex	F	M
average_speed	count	3824.000000	4084.000000
	mean	3.620257	5.871937
	std	2.162280	32.467571
	min	0.000000	0.000000
	25%	2.293500	3.183000
	50%	2.966000	5.293500
	75%	5.085000	7.037000
	max	40.066000	1888.900000
distance	count	3824.000000	4084.000000
	mean	15898.547673	27363.592483
	std	23860.361139	33537.655537
	min	0.000000	0.000000
	25%	4362.400000	7204.350000
	50%	7914.650000	16470.500000
	75%	18651.425000	37776.350000
	max	743883.000000	938421.000000

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.

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          average_speed   R-squared:                       0.496
Model:                            OLS   Adj. R-squared:                  0.496
Method:                 Least Squares   F-statistic:                     3757.
Date:                Mon, 26 Nov 2018   Prob (F-statistic):               0.00
Time:                        17:32:09   Log-Likelihood:                -9620.1
No. Observations:                3824   AIC:                         1.924e+04
Df Residuals:                    3823   BIC:                         1.925e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
distance       0.0001   1.69e-06     61.298      0.000       0.000       0.000
==============================================================================
Omnibus:                     5056.547   Durbin-Watson:                   1.114
Prob(Omnibus):                  0.000   Jarque-Bera (JB):          8433546.139
Skew:                          -6.543   Prob(JB):                         0.00
Kurtosis:                     232.693   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

Model for Female

```
                            OLS Regression Results
==============================================================================
Dep. Variable:          average_speed   R-squared:                       0.018
Model:                            OLS   Adj. R-squared:                  0.018
Method:                 Least Squares   F-statistic:                     76.65
Date:                Mon, 26 Nov 2018   Prob (F-statistic):           2.94e-18
Time:                        17:32:27   Log-Likelihood:                -20036.
No. Observations:                4084   AIC:                         4.007e+04
Df Residuals:                    4083   BIC:                         4.008e+04
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
distance       0.0001   1.18e-05      8.755      0.000    8.03e-05       0.000
==============================================================================
Omnibus:                    13078.110   Durbin-Watson:                   1.982
Prob(Omnibus):                  0.000   Jarque-Bera (JB):       1390444862.384
Skew:                          51.637   Prob(JB):                         0.00
Kurtosis:                    2859.643   Cond. No.                         1.00
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
```

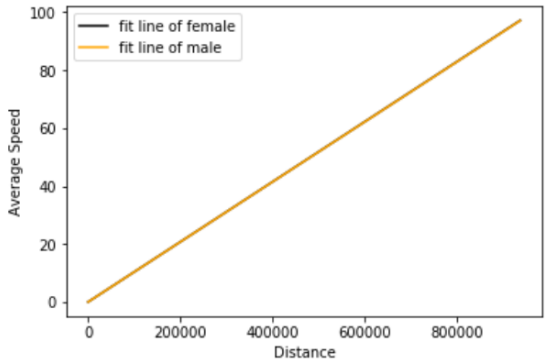
Model for Male

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.



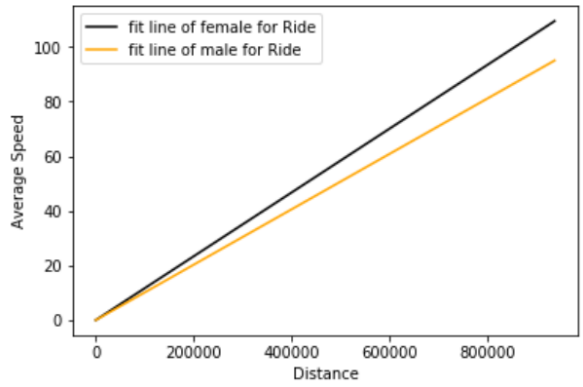
It seems that the fitted lines for them are kinds of overlapping, which means men do exercise as intensely as women.

Discussion

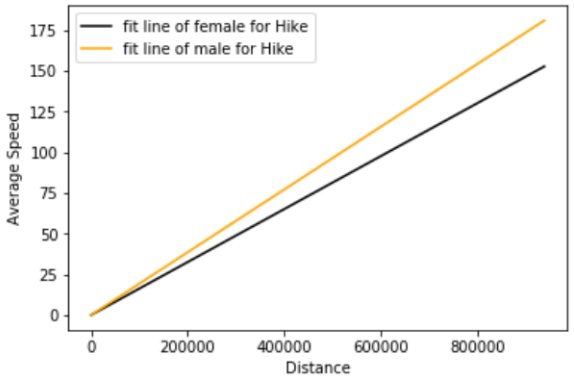
From above, we got the conclusion that there is no gender difference in exercise intensity when we are considering overall exercises. However, we are hard to say that men and women perform same level of intensity for every exercise. And, we think the reason lead to this conclusion because men do some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

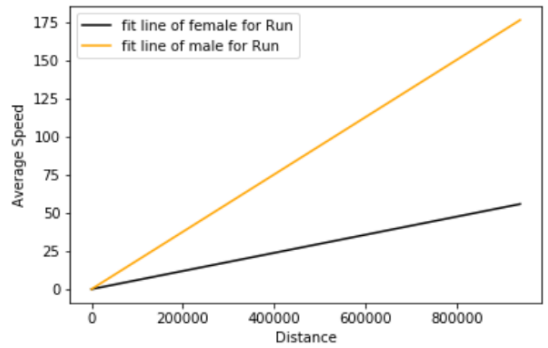
```
Model of female for Ride
distance    0.000117
dtype: float64
Model of male for Ride
distance    0.000101
dtype: float64
```



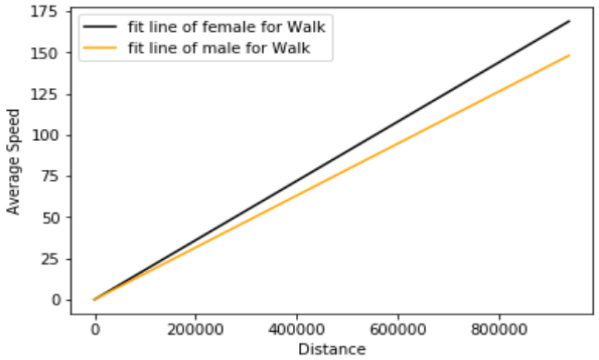

```
Model of female for Hike
distance    0.000163
dtype: float64
Model of male for Hike
distance    0.000193
dtype: float64
```



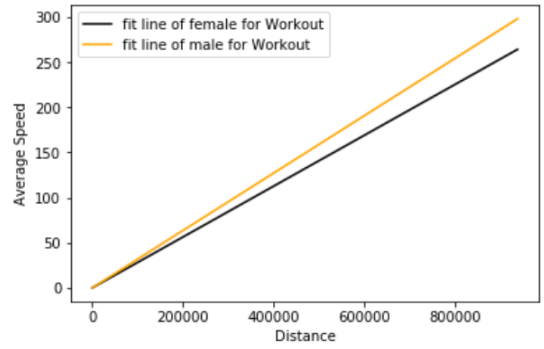

```
Model of female for Run
distance    0.00006
dtype: float64
Model of male for Run
distance    0.000188
dtype: float64
```



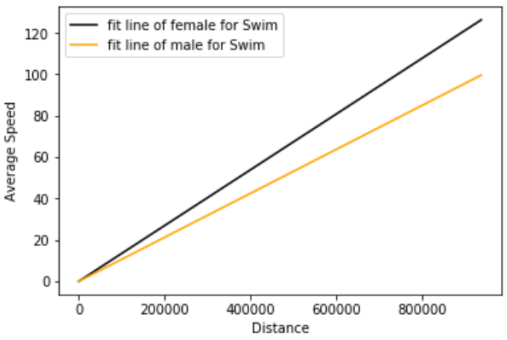

```
Model of female for Walk
distance    0.00018
dtype: float64
Model of male for Walk
distance    0.000158
dtype: float64
```



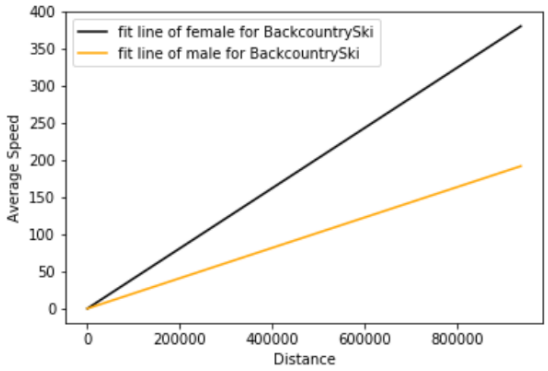

```
Model of female for Workout
distance    0.000282
dtype: float64
Model of male for Workout
distance    0.000318
dtype: float64
```



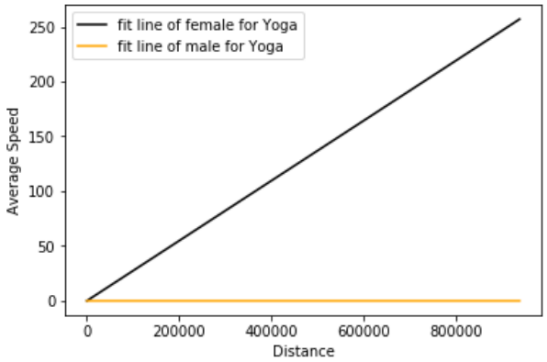

```
Model of female for Swim
distance    0.000135
dtype: float64
Model of male for Swim
distance    0.000106
dtype: float64
```



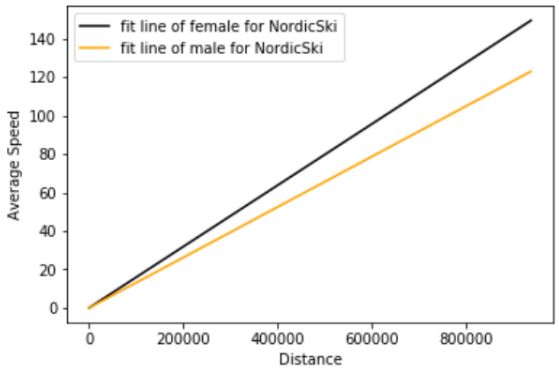
```
Model of female for BackcountrySki
distance    0.000405
dtype: float64
Model of male for BackcountrySki
distance    0.000204
dtype: float64
```



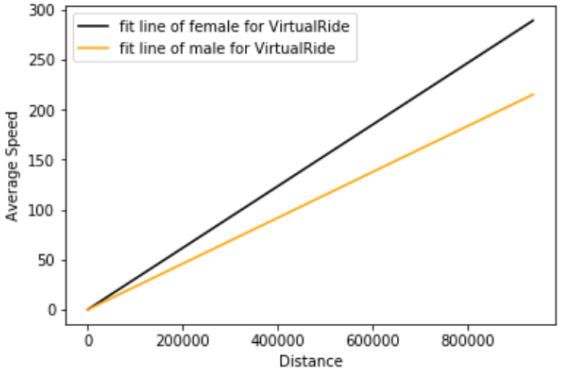

```
Model of female for Yoga
distance    0.000274
dtype: float64
Model of male for Yoga
distance    0.0
dtype: float64
```



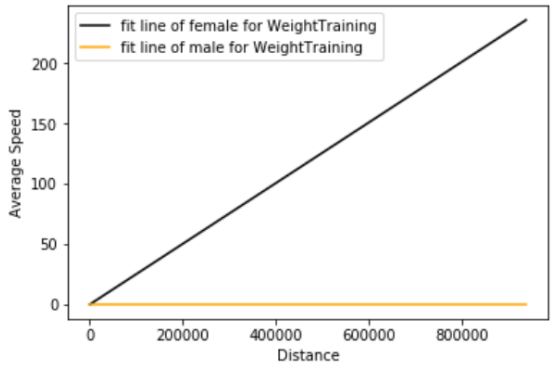

```
Model of female for NordicSki
distance    0.000159
dtype: float64
Model of male for NordicSki
distance    0.000131
dtype: float64
```




```
Model of female for VirtualRide
distance    0.000308
dtype: float64
Model of male for VirtualRide
distance    0.000229
dtype: float64
```

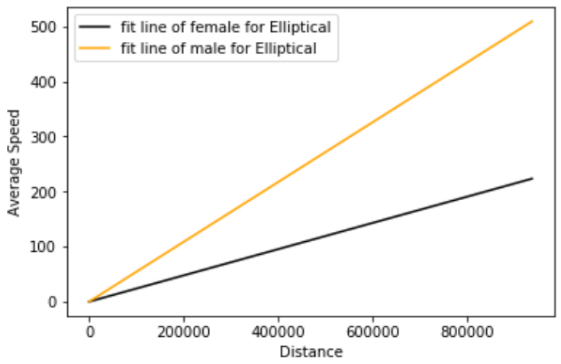



```
Model of female for WeightTraining
distance    0.000251
dtype: float64
Model of male for WeightTraining
distance    0.0
dtype: float64
```

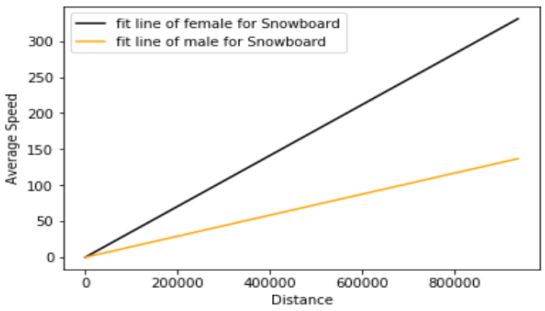



```
Model of female for Crossfit
distance    0.000182
dtype: float64
Model of male for Crossfit
distance    0.000128
dtype: float64
```

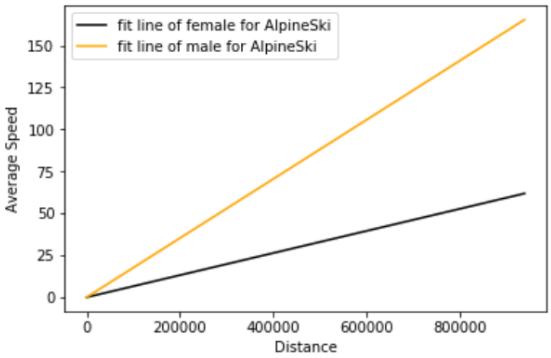
```
Model of female for Elliptical
distance    0.000238
dtype: float64
Model of male for Elliptical
distance    0.000543
dtype: float64
```



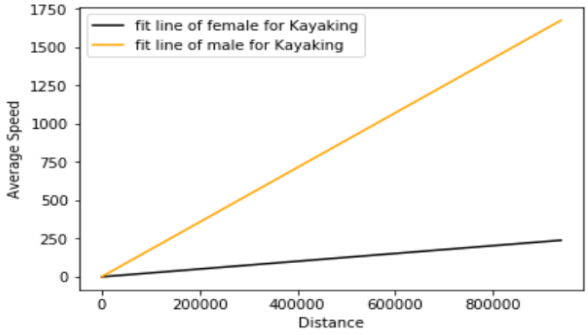

```
Model of female for Snowboard
distance    0.000353
dtype: float64
Model of male for Snowboard
distance    0.000146
dtype: float64
```




```
Model of female for AlpineSki
distance    0.000066
dtype: float64
Model of male for AlpineSki
distance    0.000176
dtype: float64
```




```
Model of female for Kayaking
distance    0.000254
dtype: float64
Model of male for Kayaking
distance    0.001783
dtype: float64
```



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

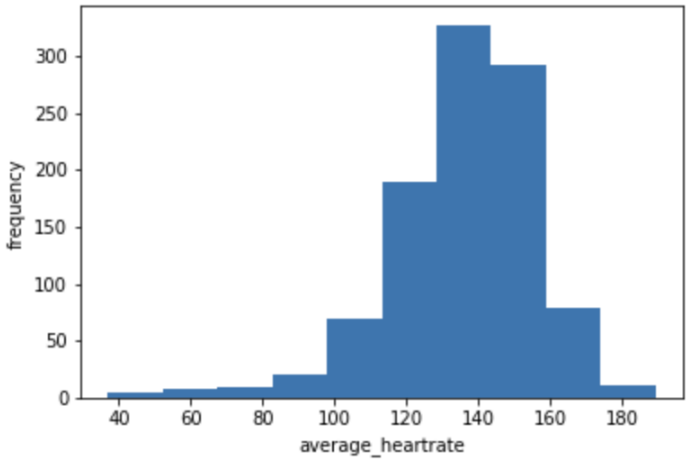
value in the columns of ‘athlete.country’ and ‘location_country’ and we would only use the observations with same value in the columns of ‘athlete.country’ and ‘location_country’. After narrowing the data, we chose three columns. They are ‘average_heartrate’, ‘total_elevation_gain’ and ‘distance’. ‘average_heartrate’ 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

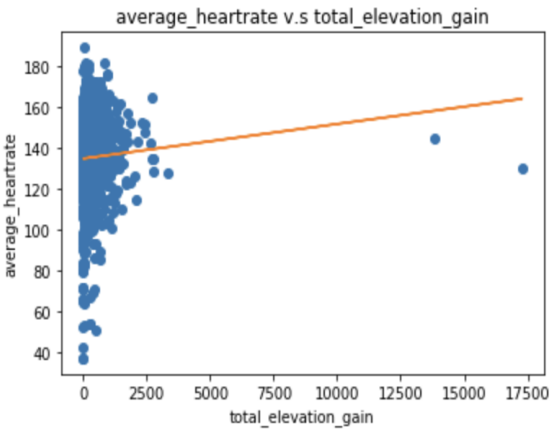
Firstly, we got the summary of this dataset.

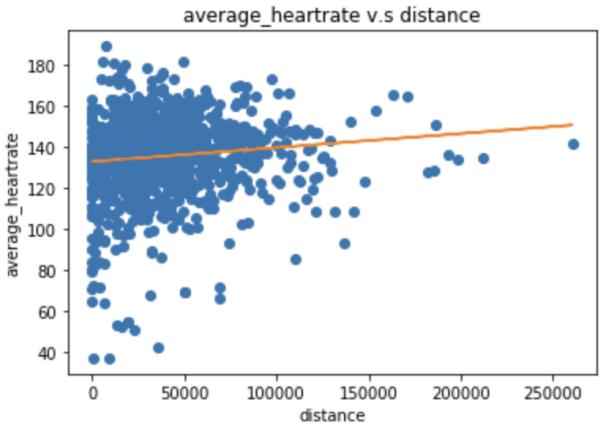
	average_heartrate	total_elevation_gain	distance
count	1010.000000	1010.000000	1010.000000
mean	135.596535	398.238812	39760.291386
std	20.249523	807.285357	31730.162093
min	37.000000	0.000000	0.000000
25%	125.325000	55.000000	19232.275000
50%	138.100000	236.400000	32476.750000
75%	148.175000	509.750000	52463.900000
max	189.200000	17281.000000	260448.000000

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 total elevation gained does.

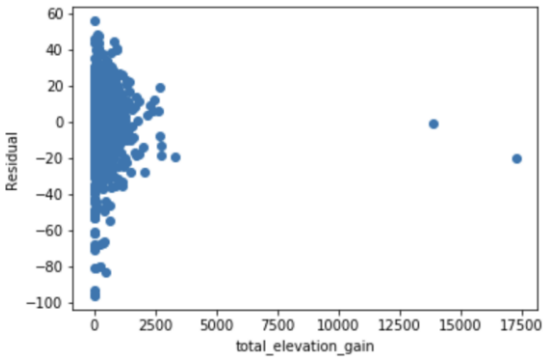
Alternative hypothesis: the total elevation ride athletes gained contributes more to their heart rate than the ride distance does.

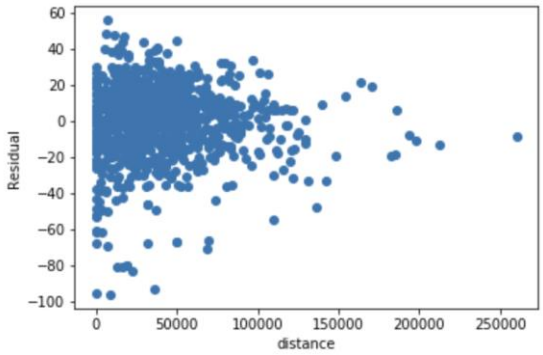
We ran a multivariable linear regression model and we got the summary below:

```
                            OLS Regression Results
==============================================================================
Dep. Variable:      average_heartrate   R-squared:                       0.013
Model:                            OLS   Adj. R-squared:                  0.011
Method:                 Least Squares   F-statistic:                     6.476
Date:                Mon, 26 Nov 2018   Prob (F-statistic):            0.00161
Time:                        21:03:26   Log-Likelihood:                -4464.4
No. Observations:                1010   AIC:                             8935.
Df Residuals:                    1007   BIC:                             8950.
Df Model:                           2
Covariance Type:            nonrobust
============================================================================================
                           coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------------------
const                  132.8102      1.018    130.449      0.000     130.812     134.808
total_elevation_gain     0.0009      0.001      1.113      0.266      -0.001       0.003
distance              6.083e-05   2.11e-05      2.882      0.004    1.94e-05       0.000
==============================================================================
Omnibus:                      204.451   Durbin-Watson:                   1.949
Prob(Omnibus):                  0.000   Jarque-Bera (JB):              510.764
Skew:                          -1.067   Prob(JB):                    1.23e-111
Kurtosis:                       5.754   Cond. No.                     8.17e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.
[2] The condition number is large, 8.17e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
```

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.

```
model.t_test("total_elevation_gain = distance")
```

```
<class 'statsmodels.stats.contrast.ContrastResults'>
                             Test for Constraints
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
c0             0.0009      0.001      1.031      0.303      -0.001       0.003
==============================================================================
```

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 total elevation gained 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 total elevation gained 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.