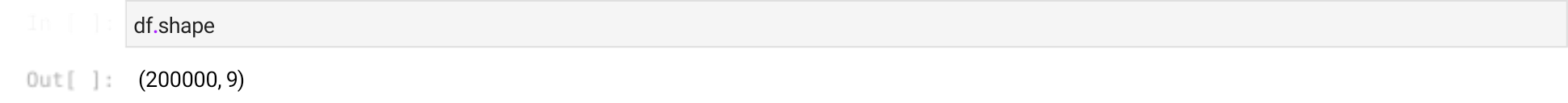


<class 'pandas.core.frame.DataFrame'> RangeIndex: 200000 entries, 0 to 199999 Data columns (total 9 columns):

# Column Non-Null Count Dtype

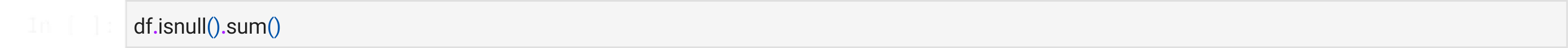
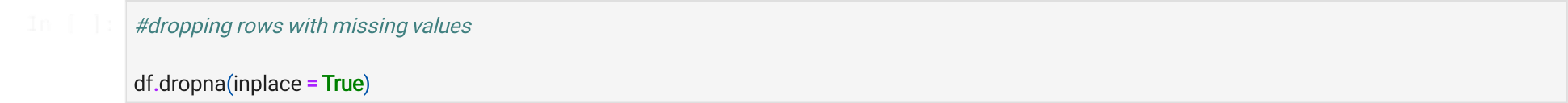


1. Unnamed: 0 200000 non-null int64
2. key 200000 non-null object
3. fare\_amount 200000 non-null float64
4. pickup\_datetime 200000 non-null object
5. pickup\_longitude 200000 non-null float64
6. pickup\_latitude 200000 non-null float64
7. dropoff\_longitude 199999 non-null float64
8. dropoff\_latitude 199999 non-null float64
9. passenger\_count 200000 non-null int64 dtypes: float64(5), int64(2), object(2) memory usage: 13.7+ MB



1. Pre-process the dataset.

| Unnamed: 0 | 0 |
| --- | --- |
| key | 0 |
| fare\_amount | 0 |
| pickup\_datetime | 0 |
| pickup\_longitude | 0 |
| pickup\_latitude | 0 |
| dropoff\_longitude | 1 |
| dropoff\_latitude | 1 |
| passenger\_count | 0 |
| dtype: int64 |  |



Unnamed: 0 0

key 0

fare\_amount 0

pickup\_datetime 0

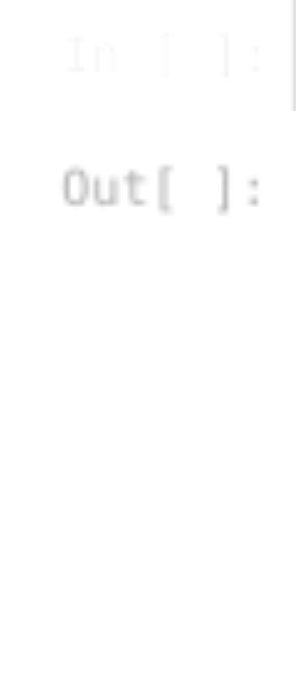
pickup\_longitude 0

pickup\_latitude 0

dropoff\_longitude 0

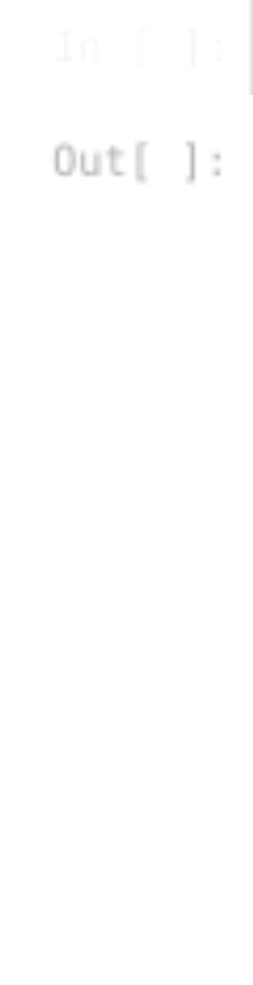
dropoff\_latitude 0

passenger\_count 0

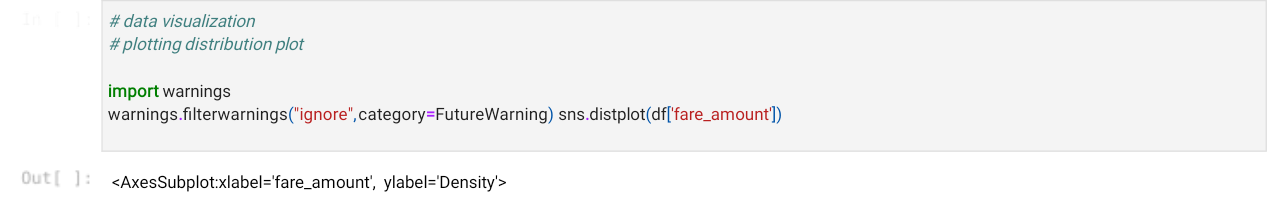
dtype: int64

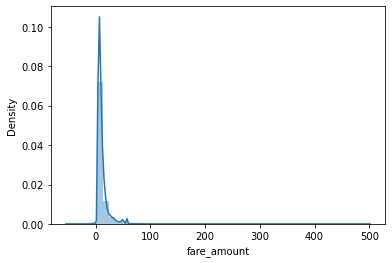


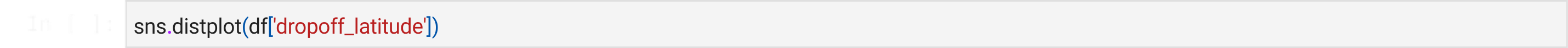
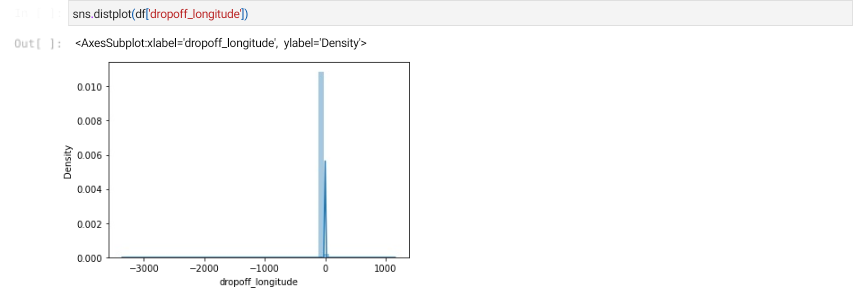
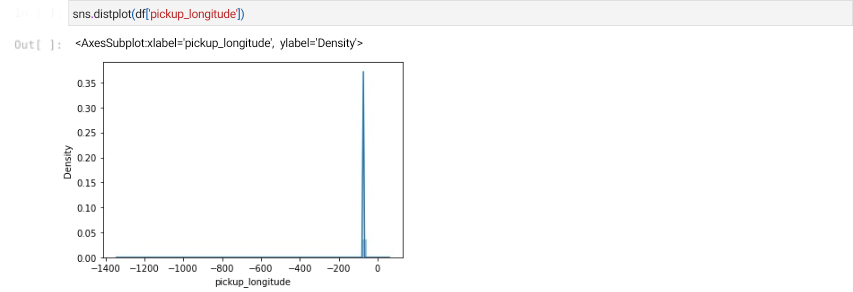
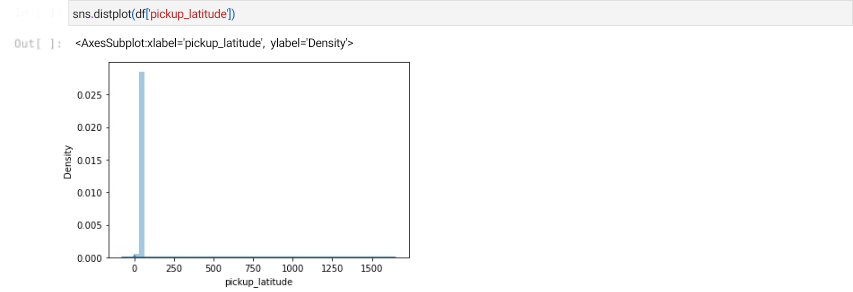
| df**.**dtypes |  | | | | | |
| --- | --- | --- | --- | --- | --- | --- |
| fare\_amount pickup\_datetime pickup\_longitude pickup\_latitude dropoff\_longitude dropoff\_latitude passenger\_count dtype: object | float64 object float64 float64 float64 float64 int64 |  |  |  |  |  |
| df**.**describe() |  |  |  |  |  |  |
| **fare\_amount** | **pickup\_longitude** | **pickup\_latitude** | **dropoff\_longitude** | **dropoff\_latitude** | **passenger\_count** |  |
| **count** 199999.000000 | 199999.000000 | 199999.000000 | 199999.000000 | 199999.000000 | 199999.000000 |  |
| **mean** 11.359892 | -72.527631 | 39.935881 | -72.525292 | 39.923890 | 1.684543 |  |
| **std** 9.901760 | 11.437815 | 7.720558 | 13.117408 | 6.794829 | 1.385995 |  |
| **min** -52.000000 | -1340.648410 | -74.015515 | -3356.666300 | -881.985513 | 0.000000 |  |
| **25%** 6.000000 | -73.992065 | 40.734796 | -73.991407 | 40.733823 | 1.000000 |  |
| **50%** 8.500000 | -73.981823 | 40.752592 | -73.980093 | 40.753042 | 1.000000 |  |
| **75%** 12.500000 | -73.967154 | 40.767158 | -73.963658 | 40.768001 | 2.000000 |  |
| **max** 499.000000 | 57.418457 | 1644.421482 | 1153.572603 | 872.697628 | 208.000000 |  |

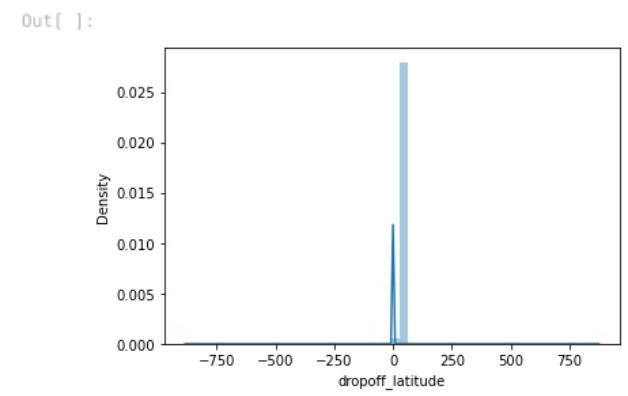
1. Identify outliers.

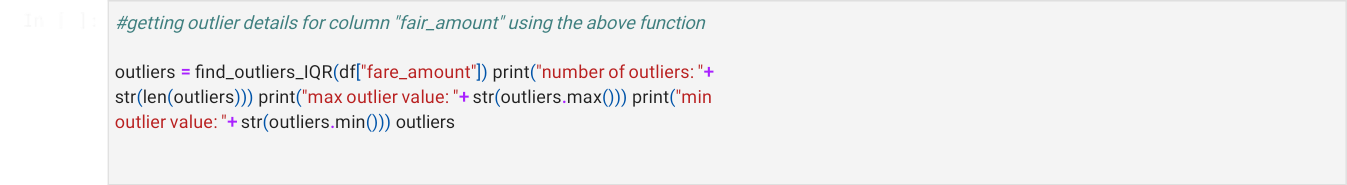
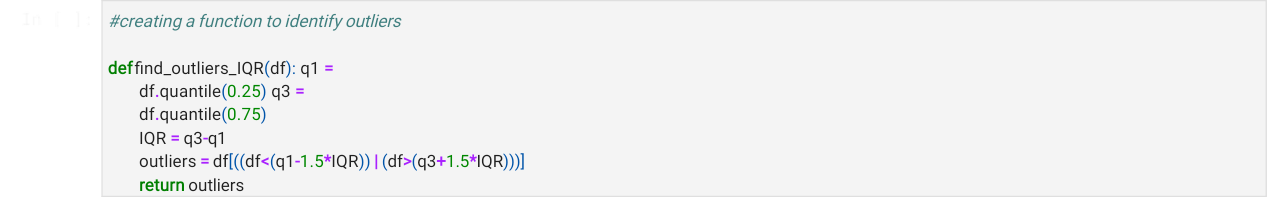
OUTLIER: An object that deviates significantly from the rest of the objects.





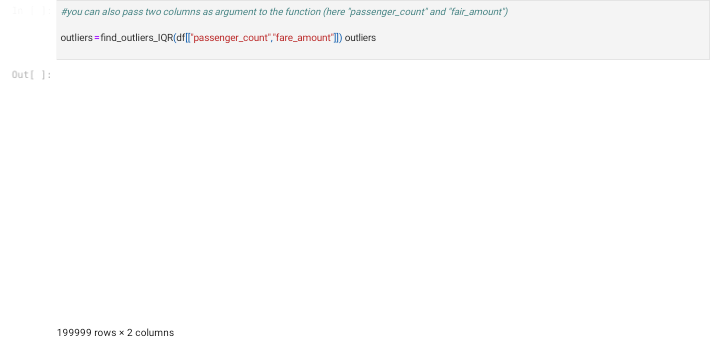


<AxesSubplot:xlabel='dropoff\_latitude', ylabel='Density'>

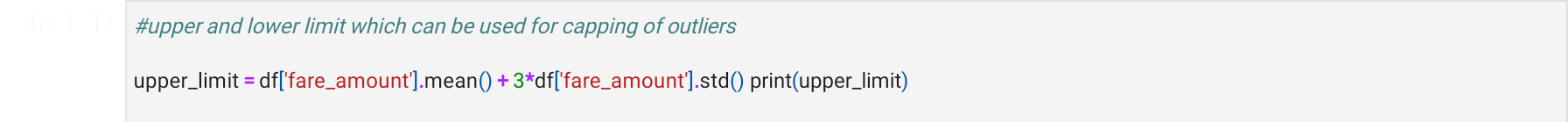


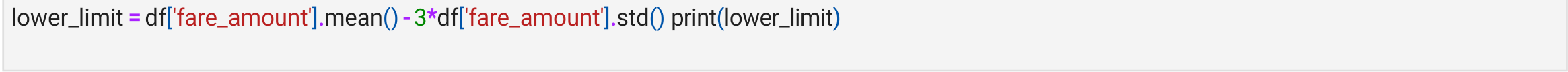
number of outliers: 17166 max outlier value: 499.0 min outlier value: -52.0

| 6 | 24.50 |
| --- | --- |
| 30 | 25.70 |
| 34 | 39.50 |
| 39 | 29.00 |
| 48 | 56.80 |
|  | ... |
| 199976 | 49.70 |
| 199977 | 43.50 |
| 199982 | 57.33 |
| 199985 | 24.00 |
| 199997 | 30.90 |

Name: fare\_amount, Length: 17166, dtype: float64

|  | **passenger\_count** | **fare\_amount** |
| --- | --- | --- |
| **0** | NaN | NaN |
| **1** | NaN | NaN |
| **2** | NaN | NaN |
| **3** | NaN | NaN |
| **4** | 5.0 | NaN |
| **...** | ... | ... |
| **199995** | NaN | NaN |
| **199996** | NaN | NaN |
| **199997** | NaN | 30.9 |
| **199998** | NaN | NaN |
| **199999** | NaN | NaN |

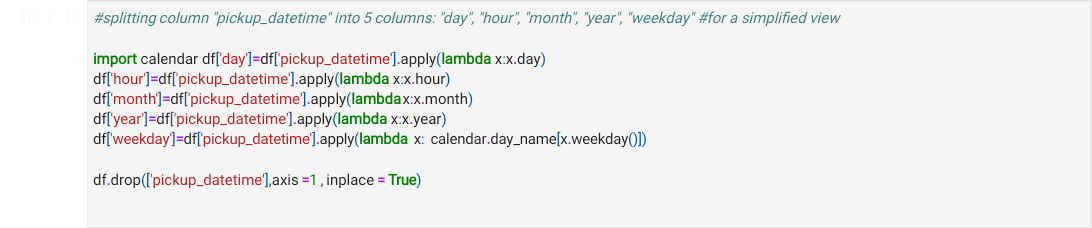
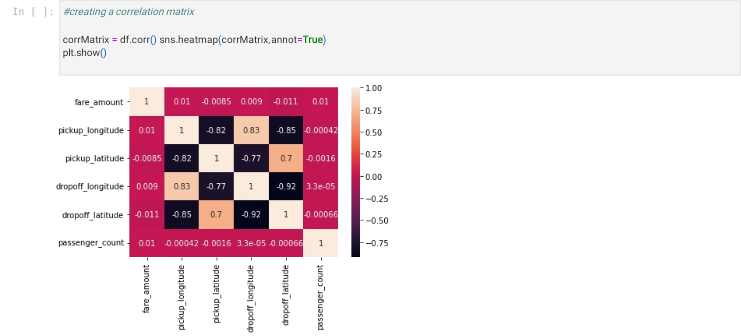




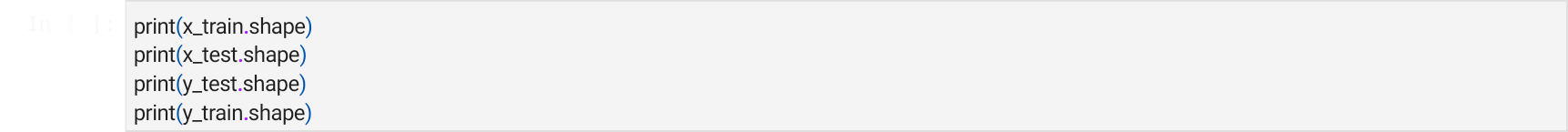
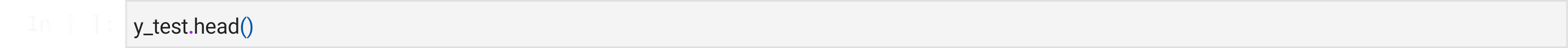
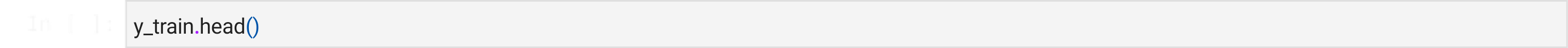
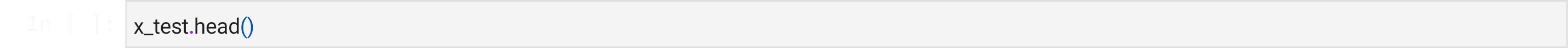
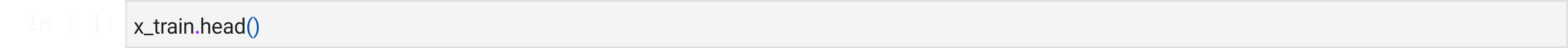
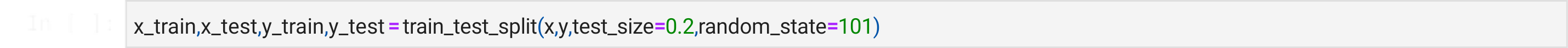
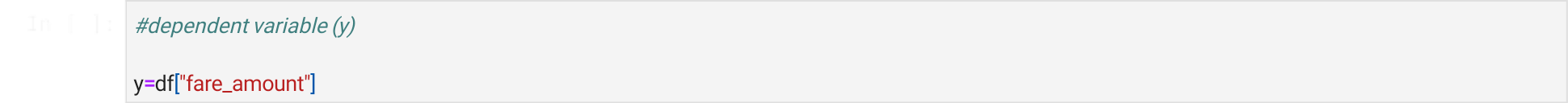
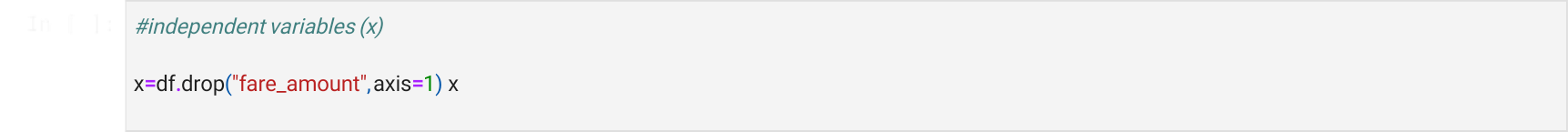
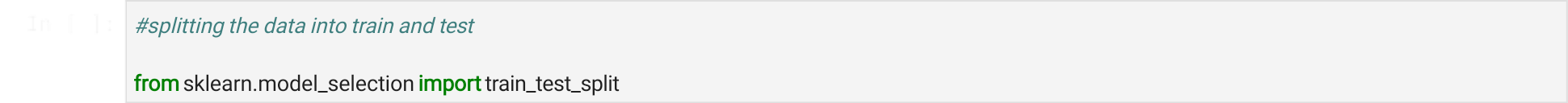
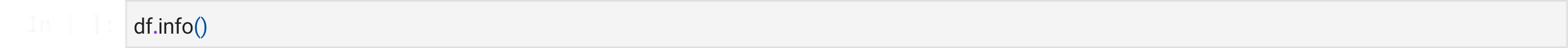
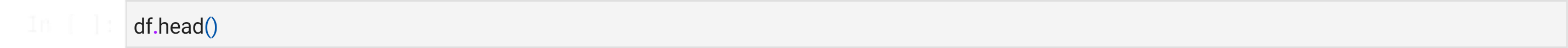
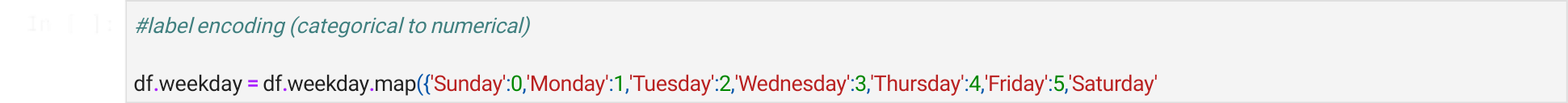
41.06517154774204

-18.3453884488253

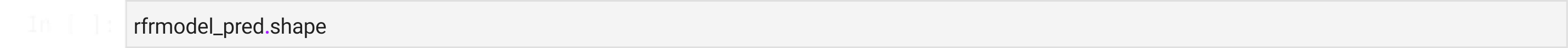
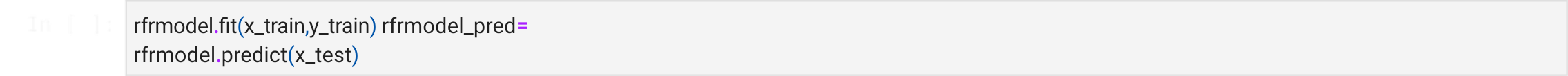
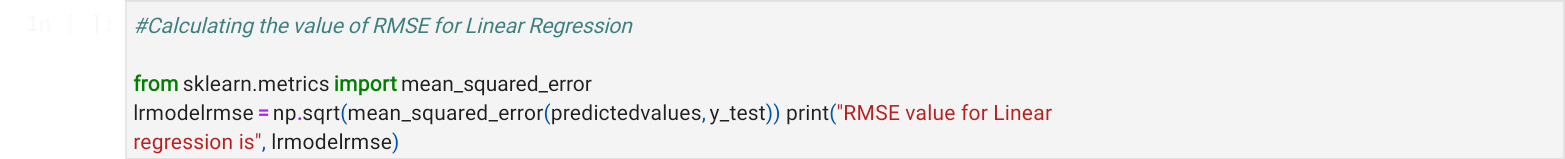
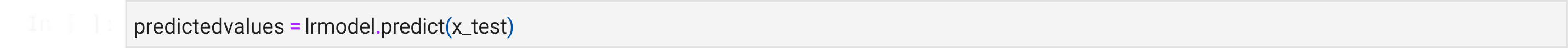
1. Check the correlation.







1. Implement linear regression and random forest regression models.
2. Evaluate the models and compare their respective scores like R2, RMSE, etc.



Predict the price of the Uber ride

