ECE 219 Project 4: Regression Analysis

Hongyi Chen (705186099), Zichun Chai (505625207), Chuang Yu (305629107)

We were working with 3 datasets which were Bike Sharing Dataset, Suicide Rates Overview 1985 to 2016, and Video Transcoding Time Dataset. The Bike Sharing Dataset provides a count number of rental bikes based on some timing and environmental conditions. We performed all the analyses based on 3 different target variables: casual: count of casual users, registered: count of registered users, and cnt: count of total rental bikes including both casual and registered.

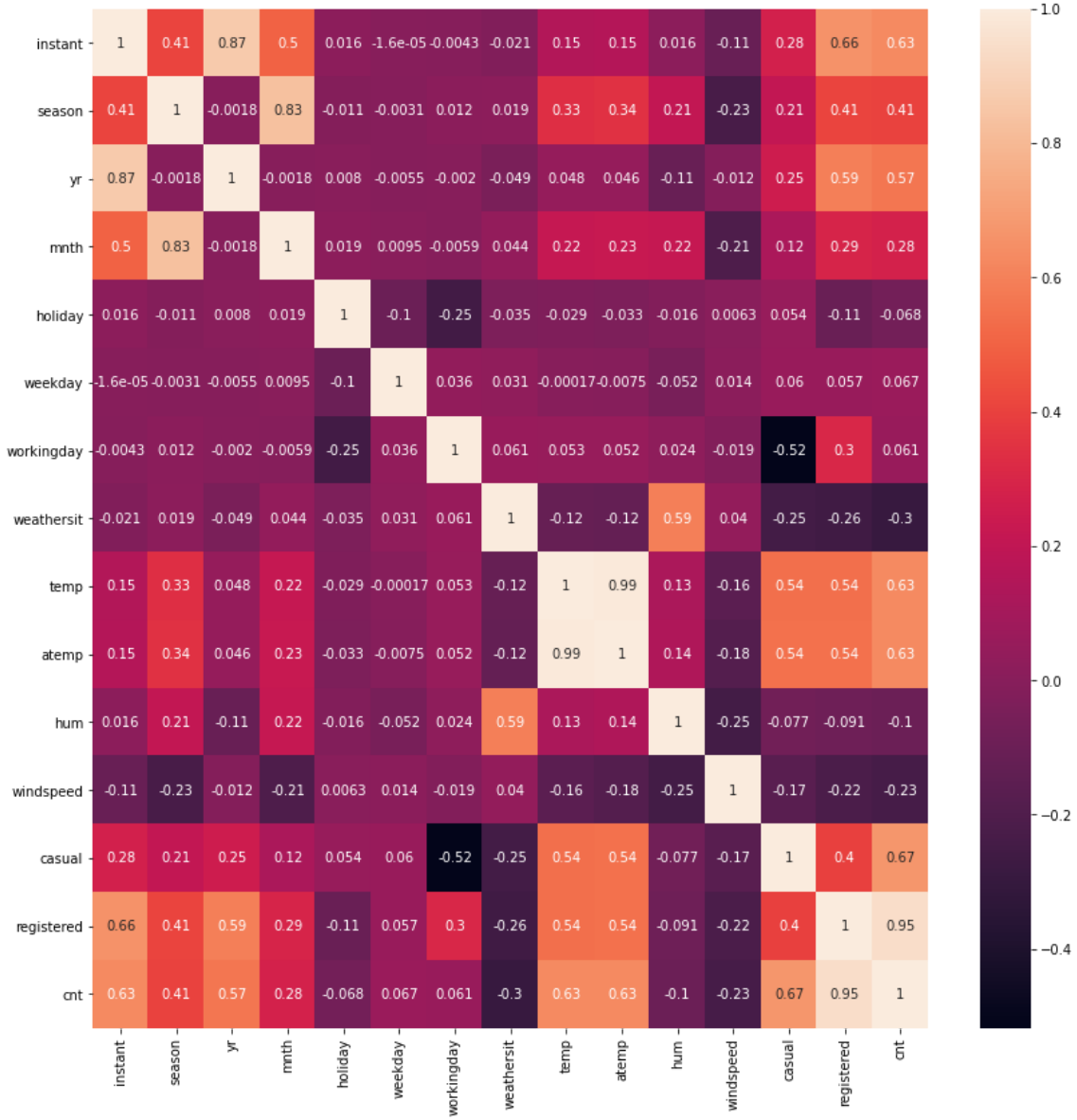
The second dataset is called Suicide Rates Overview 1985 to 2016. It is a compiled dataset that pulled from four other datasets linked by time and place, and was built to nd signals correlated to increased suicide rates among different cohorts globally, across the socioeconomic spectrum. And the target variables are suicides_no and suicides/100k pop.

The last dataset is called Video Transcoding Time Dataset. It includes input and output video characteristics along with their time taken for different valid transcodings. And the target variable is transcoding time, which is the last attribute “utime” in the data file.

Question 1:

We have plotted heatmaps of Pearson correlation matrix of dataset columns for all 3 datasets. The matrices are attached below:

Bike:



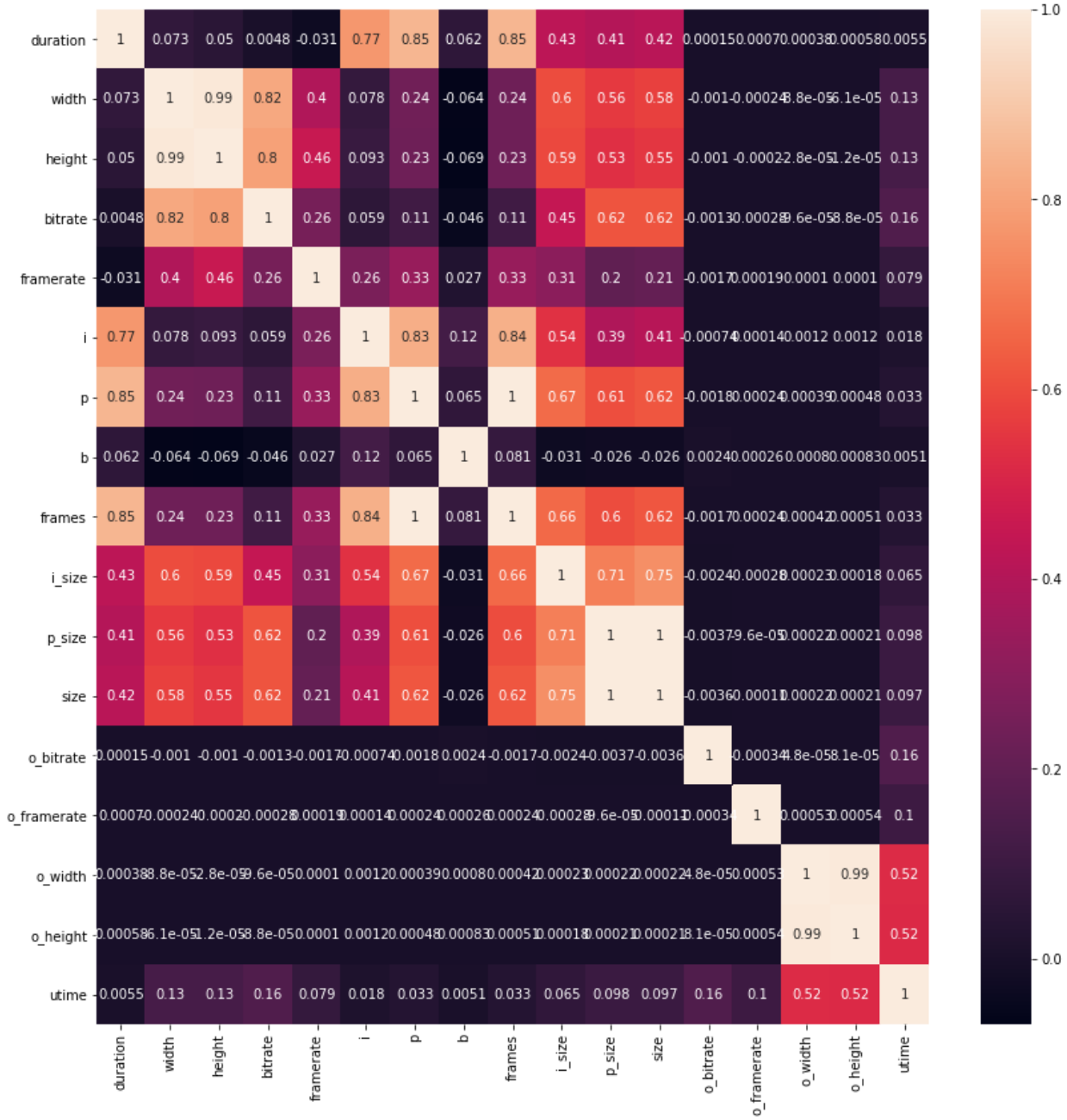
From the graph we could see that features temp and atemp have the highest correlation with the target variable cnt. It shows that temperature and feeling temperature are very important for predicting the count of total rental bikes including both casual and registered.

Suicide:


year
suicides_no
population
suicides/100k pop
gdp_for_year ($)
gdp_per_capita ($)
1
-0.0045
0.0089
-0.039
0.095
0.34
-0.0045
1
0.62
0.31
0.43
0.061
0.0089
0.62
1
0.0083
0.71
0.082
-0.039
0.31
0.0083
1
0.025
0.0018
0.095
0.43
0.71
0.025
1
0.3
0.34
0.061
0.082
0.0018
0.3
1
1.0
0.8
0.6
0.4
0.2
0.0

From the heatmap, we observed that population has the highest correlation with the target variable suicide_no, and gdp_for_year is highly correlated with another target variable suicides/100k pop.

Video:

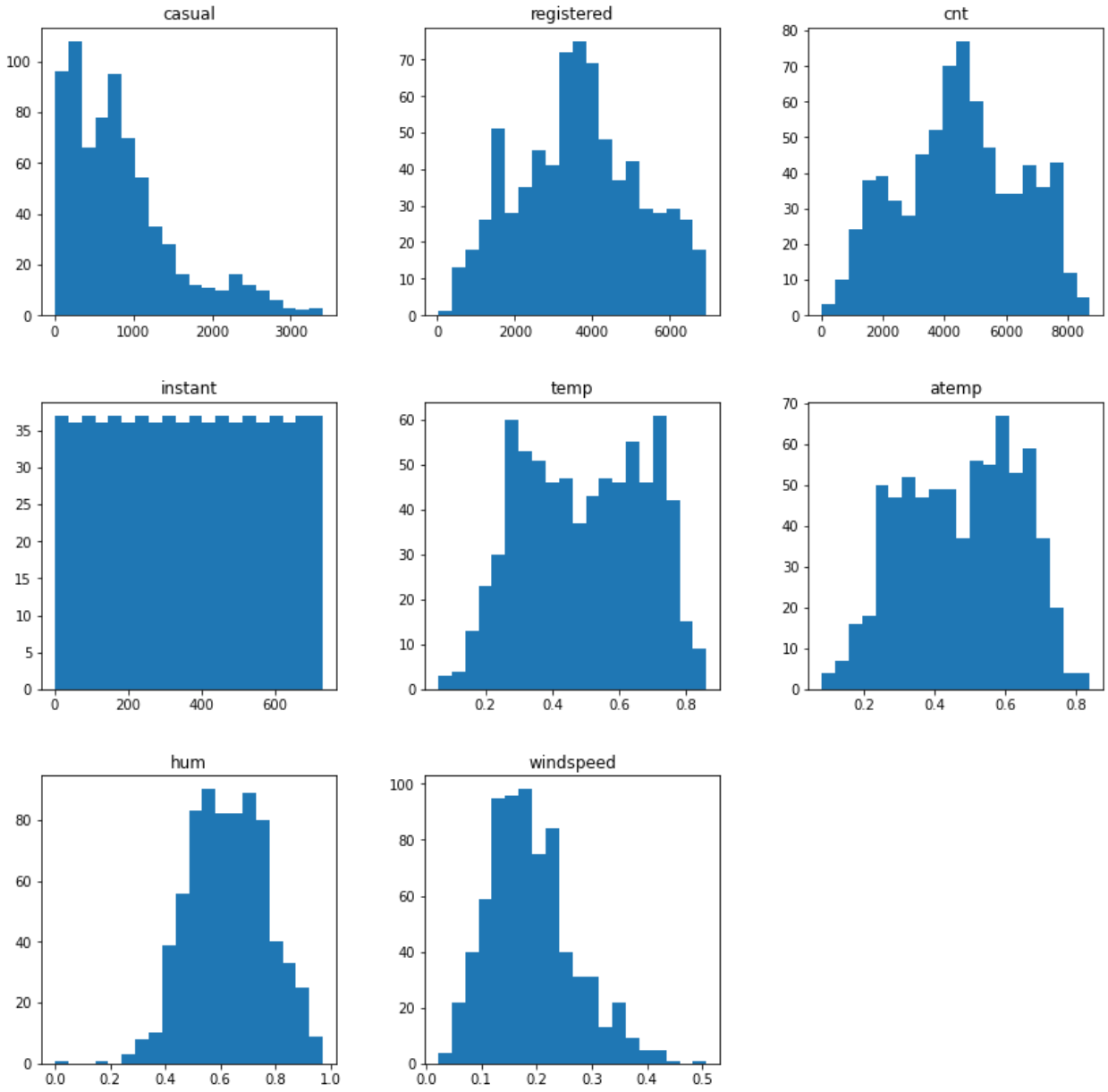


From the heatmap we could see that features o_width and o_height have the highest correlation with the target variable utime. It shows that output width and height in pixels used for transcoding are very important for predicting the total transcoding time for transcoding.

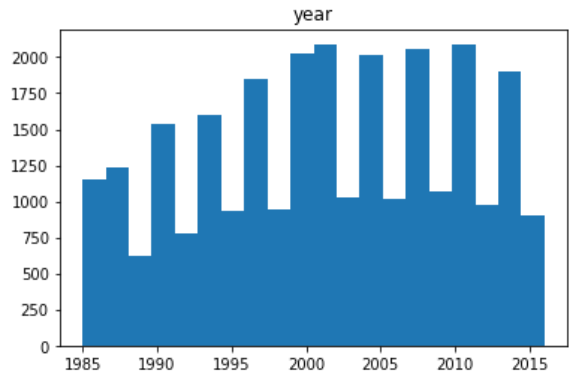
Question 2:

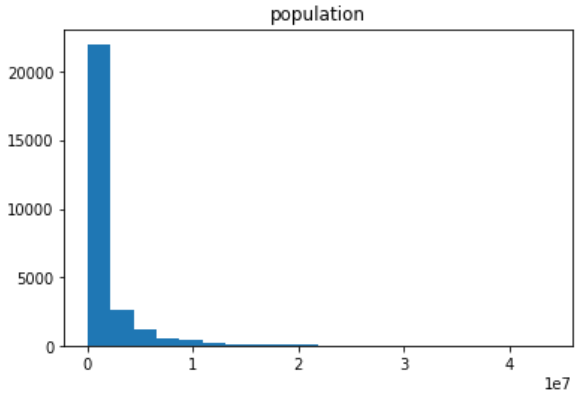
We have plotted the histograms of numerical features for all the 3 datasets and the results are shown below. Also, if the distribution of a feature has high skewness, we could possibly remove all of the observations that make it "skewed". And we could also perform data transformations to reduce skewness like logarithmic, squares, cubes, high powers, and reciprocals transformations. Standardization is another common way of reducing the skewness.

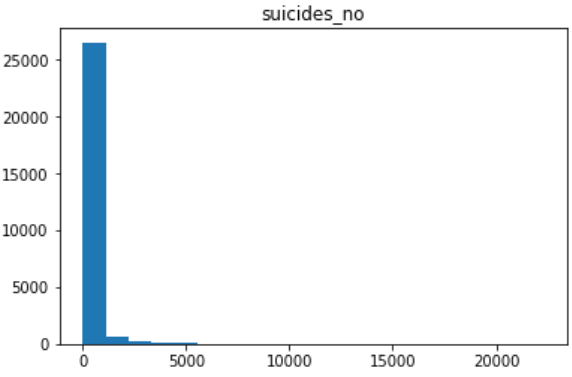
Bike:

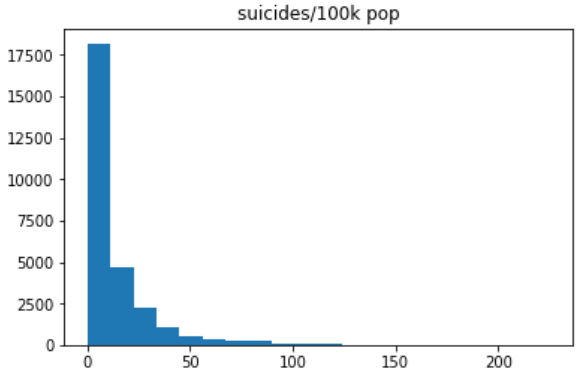


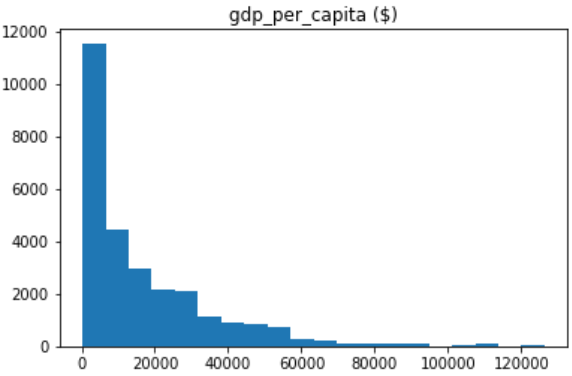
Suicide:

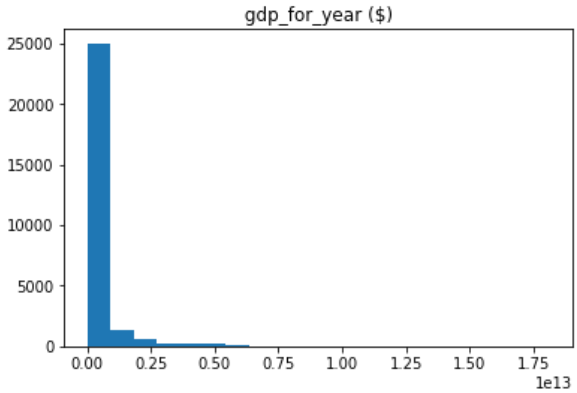

year
2000
1750
1500
1250
1000
750
500
250
0
1985
1990
1995
2000
2005
2010
2015


population
20000
15000
10000
5000
0
0
1
2
3
4
1e7

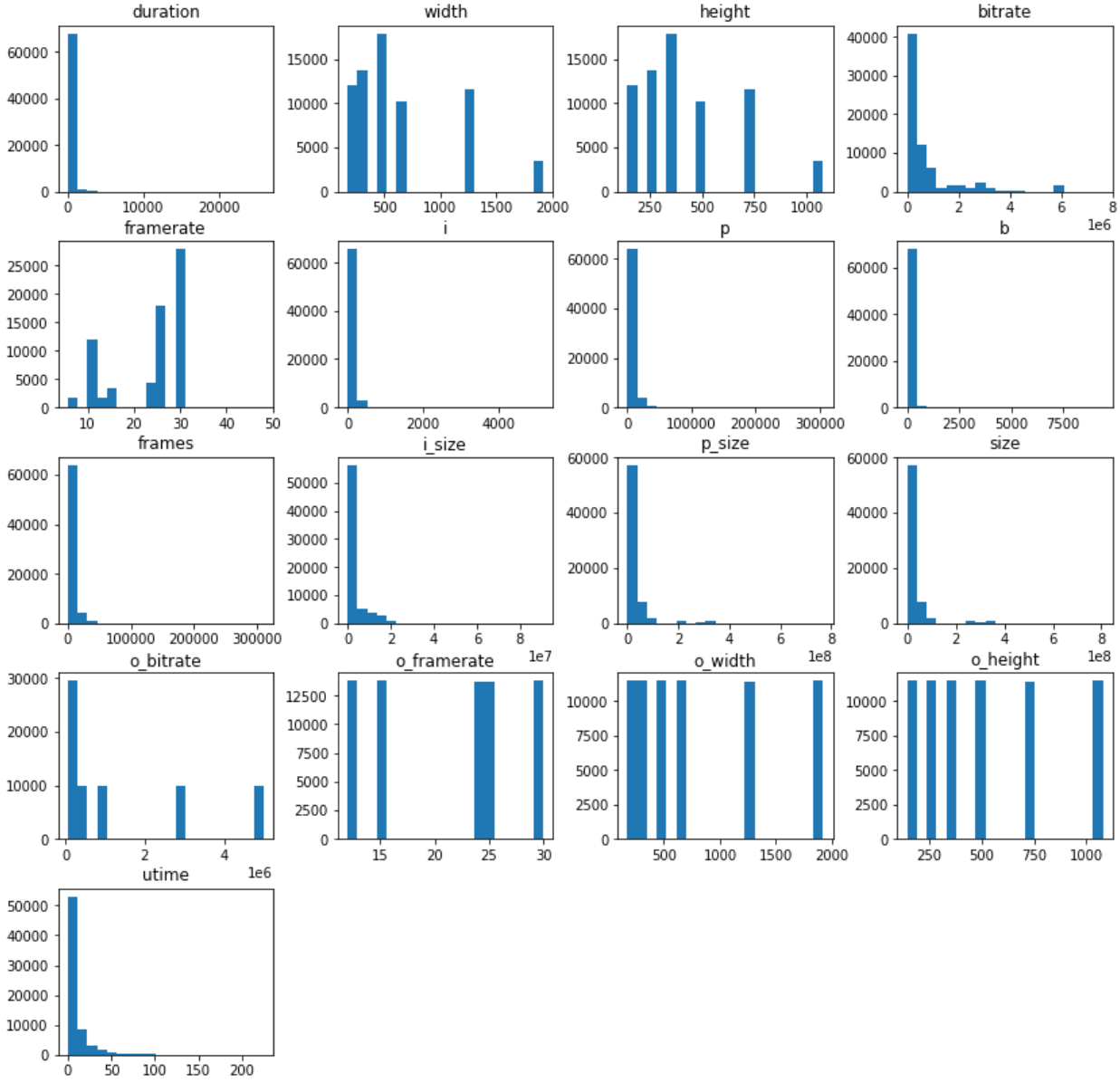

suicides_no
25000
20000
15000
10000
5000
0
0
5000
10000
15000
20000


suicides/100k pop
17500
15000
12500
10000
7500
5000
2500
0
0
50
100
150
200


gdp_per_capita ($)
12000
10000
8000
6000
4000
2000
0
0
20000
40000
60000
80000
100000
120000


gdp_for_year ($)
25000
20000
15000
10000
5000
0
0.00
0.25
0.50
0.75
1.00
1.25
1.50
1.75
1e13

Video:

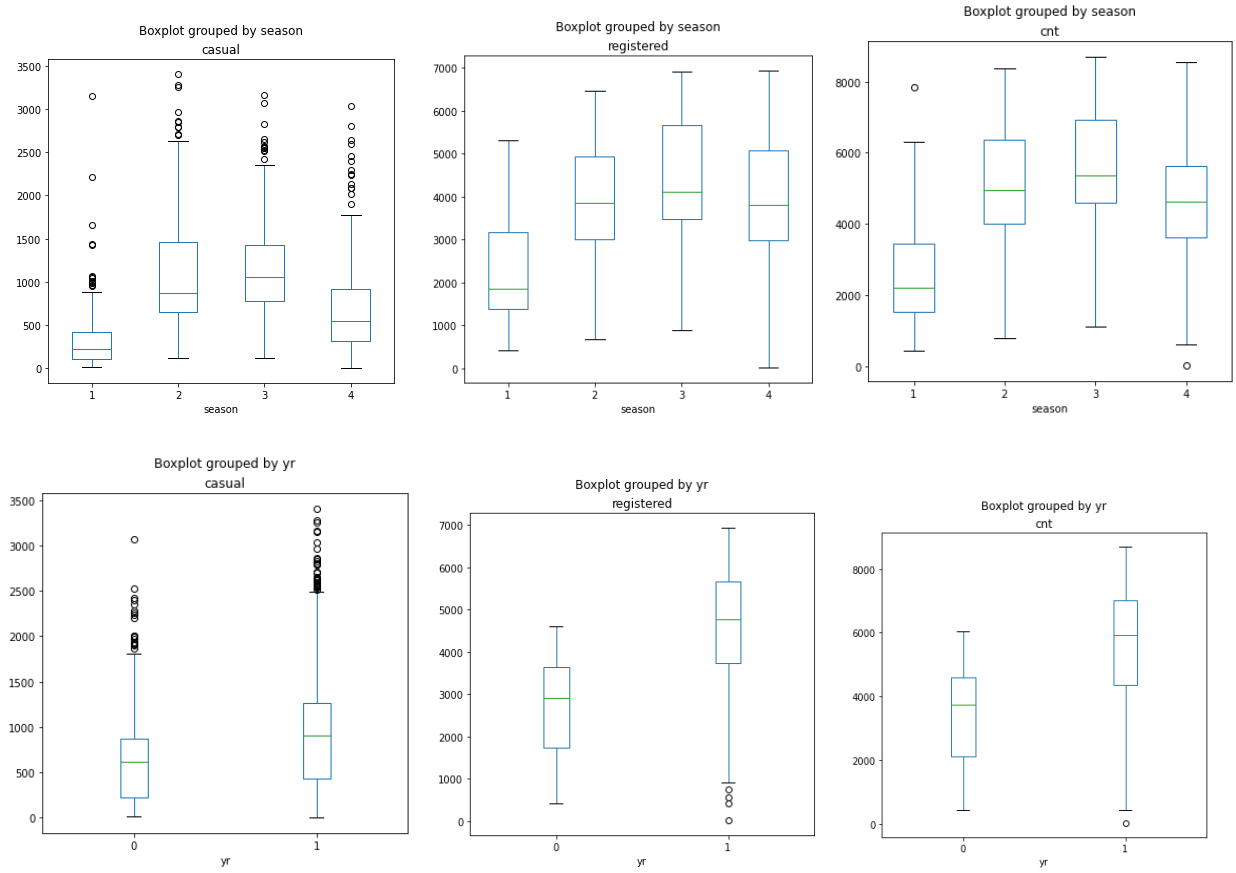

duration
width
height
bitrate
framerate
i
p
b
frames
i_size
p_size
size
o_bitrate
o_framerate
o_width
o_height
utime
1e6
1e7
1e8

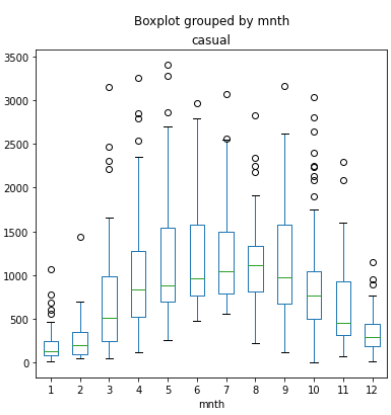
Question 3:

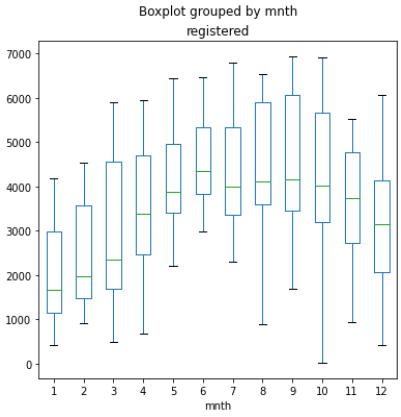
In this problem we have first plotted the box plots of categorical features vs target variable and then inspected them. Bike dataset has categorical features 'season', 'yr', 'mnth', 'holiday', 'weekday', 'workingday', and 'weathersit'. Suicide dataset has categorical features 'country', 'sex', 'age', and 'generation'. Video dataset has categorical features 'codec', and 'o_codec'.

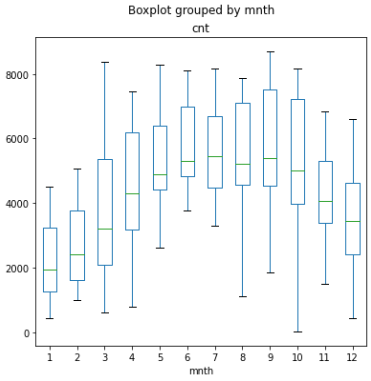
From bike's box plots, we can see that people are more willing to rent bikes during summer and fall and registered people are more likely to rent bikes during work days when casual people are more likely to rent bikes during weekend.

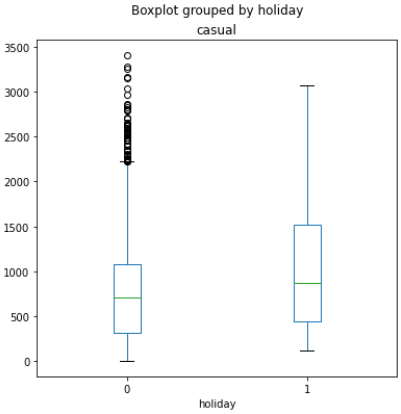
Bike:

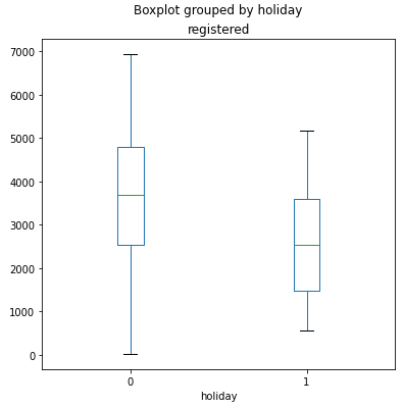


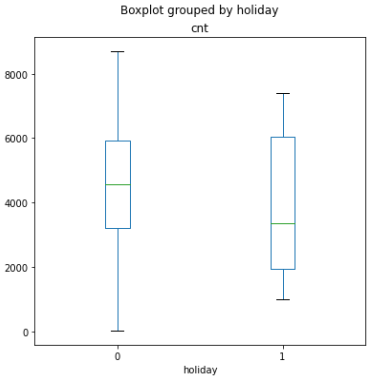

Boxplot grouped by mnth
casual
mnth

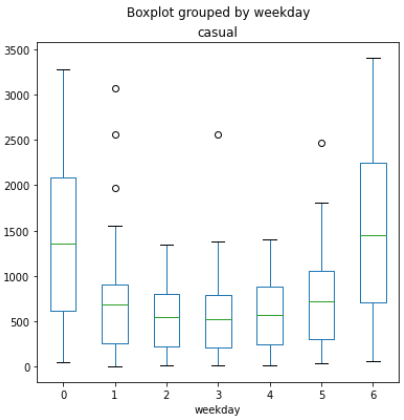

Boxplot grouped by mnth
registered
mnth

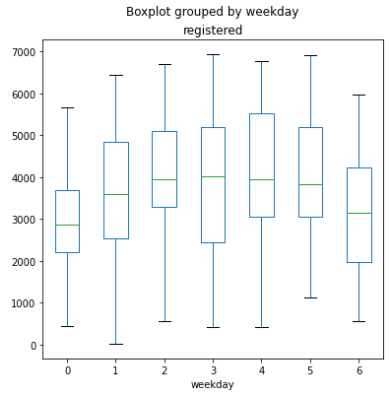

Boxplot grouped by mnth
cnt
mnth

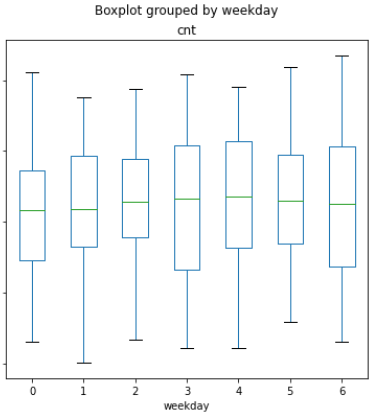

Boxplot grouped by holiday
casual
holiday

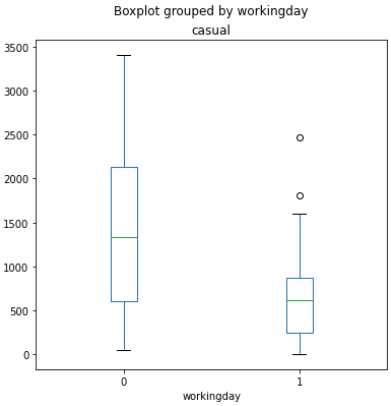

Boxplot grouped by holiday
registered
holiday

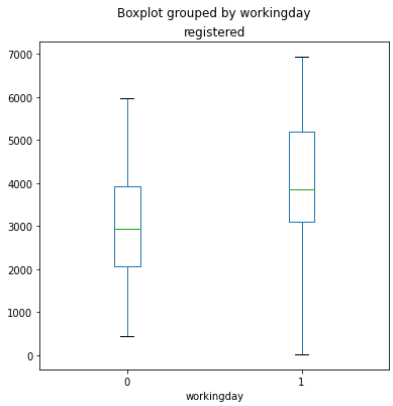

Boxplot grouped by holiday
cnt
holiday

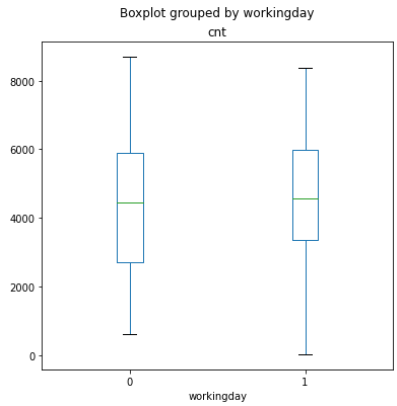

Boxplot grouped by weekday
casual
weekday

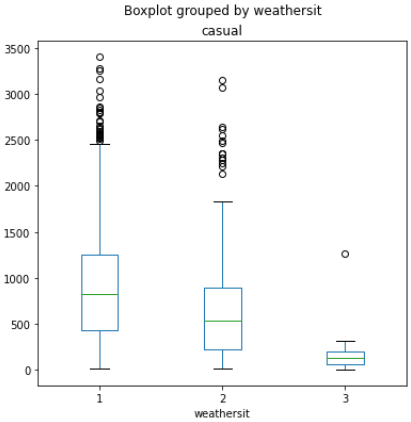

Boxplot grouped by weekday
registered
weekday

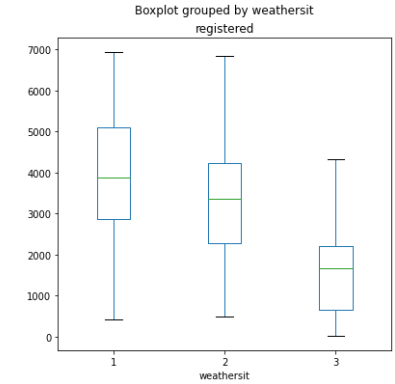

Boxplot grouped by weekday
cnt
weekday

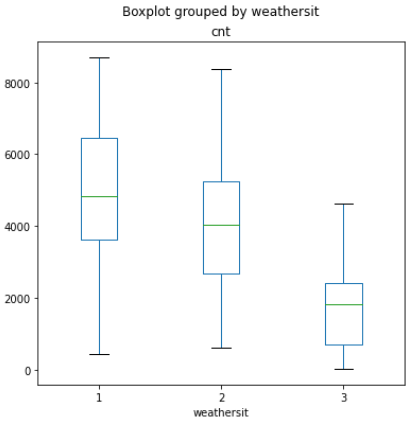

Boxplot grouped by workingday
casual
3500
3000
2500
2000
1500
1000
500
0
0
1
workingday


Boxplot grouped by workingday
registered
7000
6000
5000
4000
3000
2000
1000
0
0
1
workingday


Boxplot grouped by workingday
cnt
8000
6000
4000
2000
0
0
1
workingday

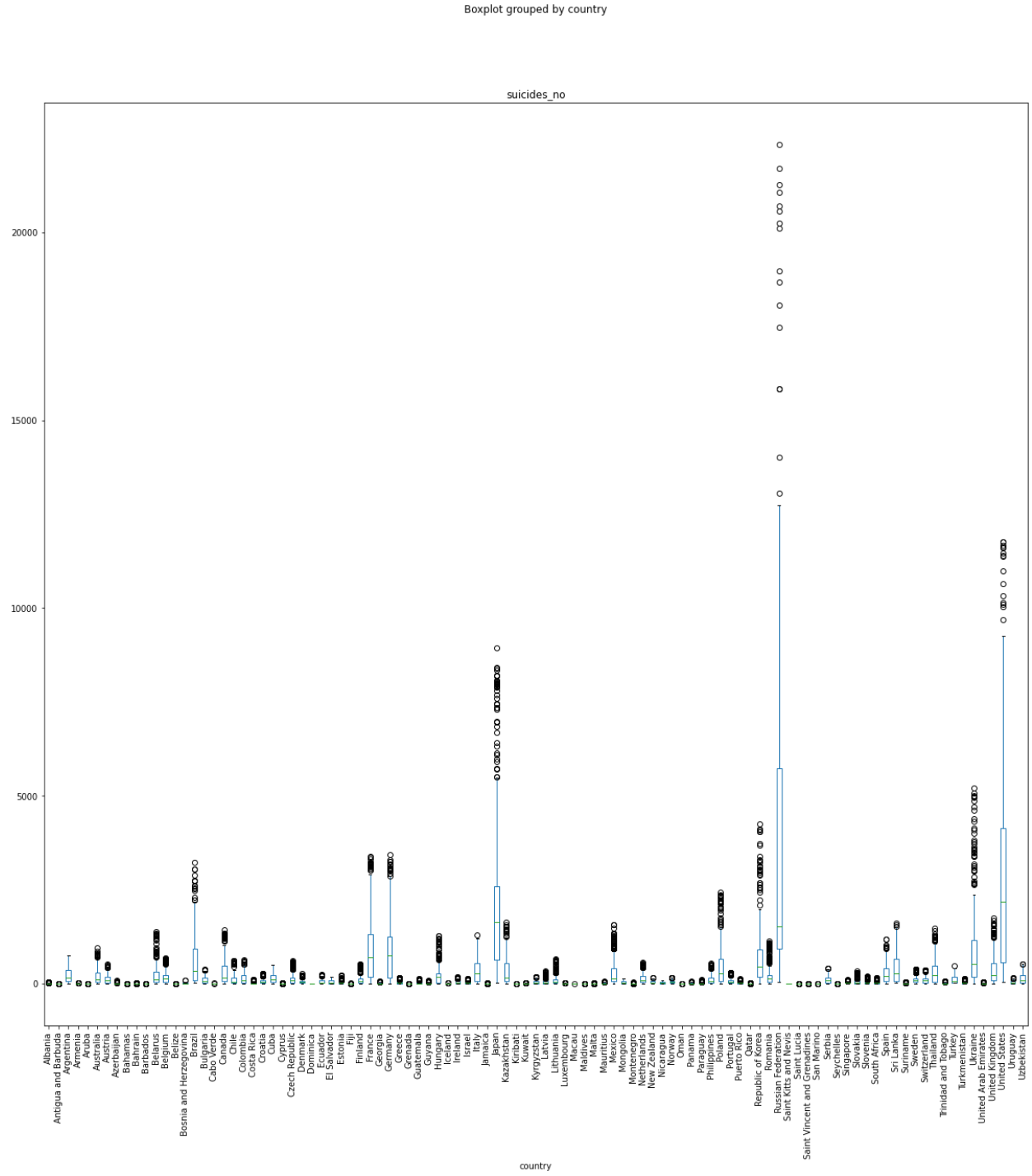

Boxplot grouped by weathersit
casual
3500
3000
2500
2000
1500
1000
500
0
1
2
3
weathersit

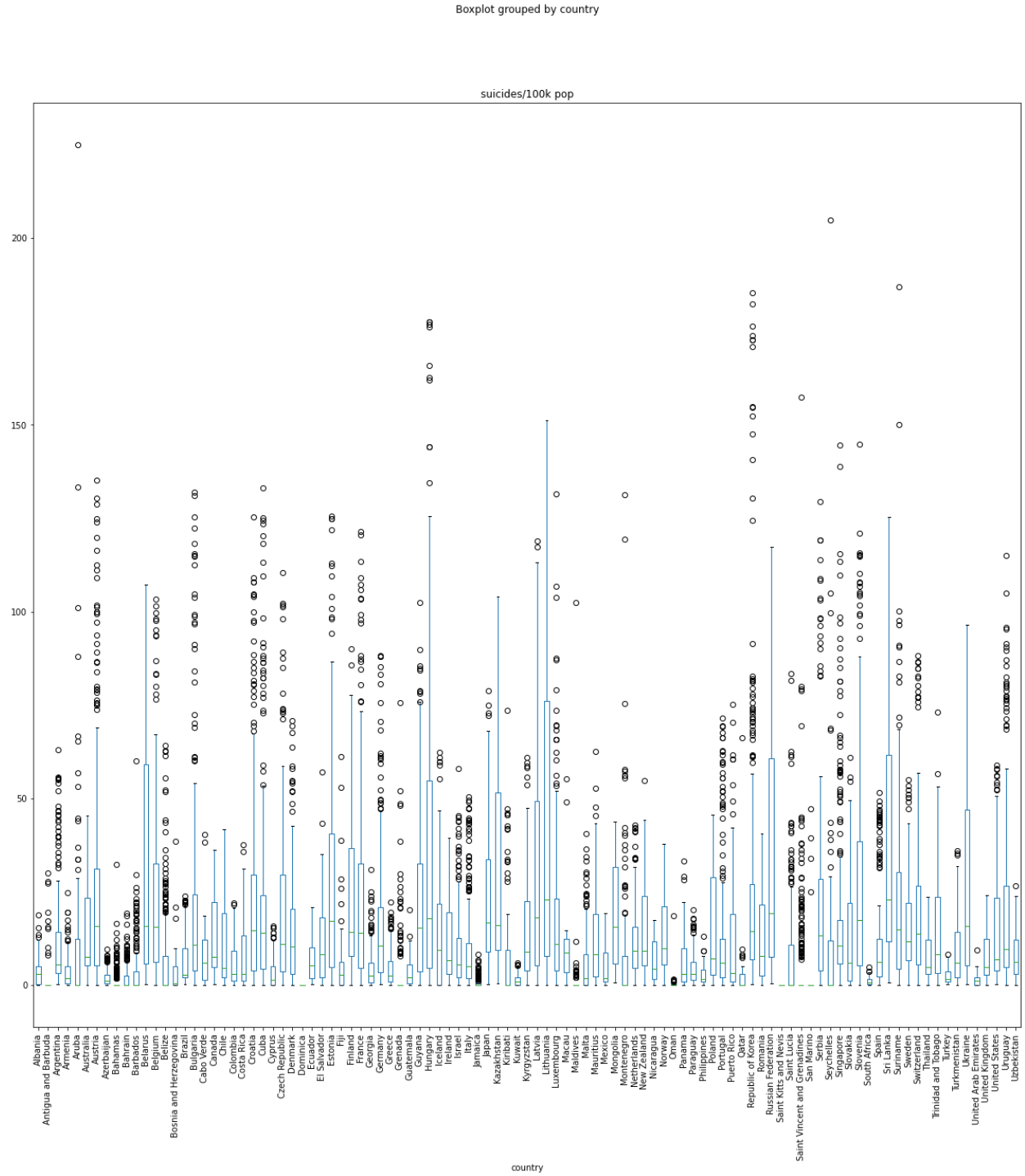

Boxplot grouped by weathersit
registered
7000
6000
5000
4000
3000
2000
1000
0
1
2
3
weathersit

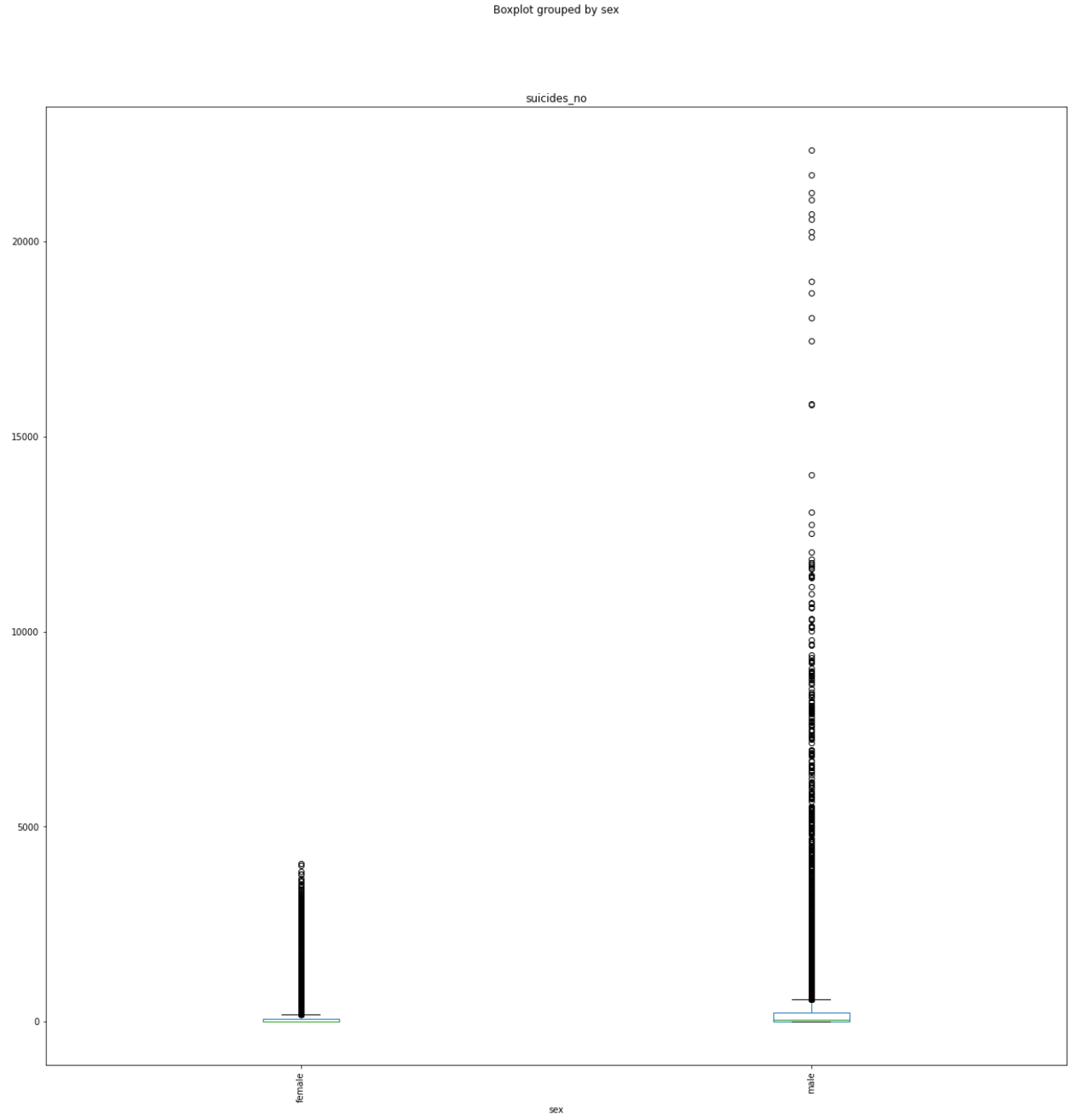

Boxplot grouped by weathersit
cnt
8000
6000
4000
2000
0
1
2
3
weathersit

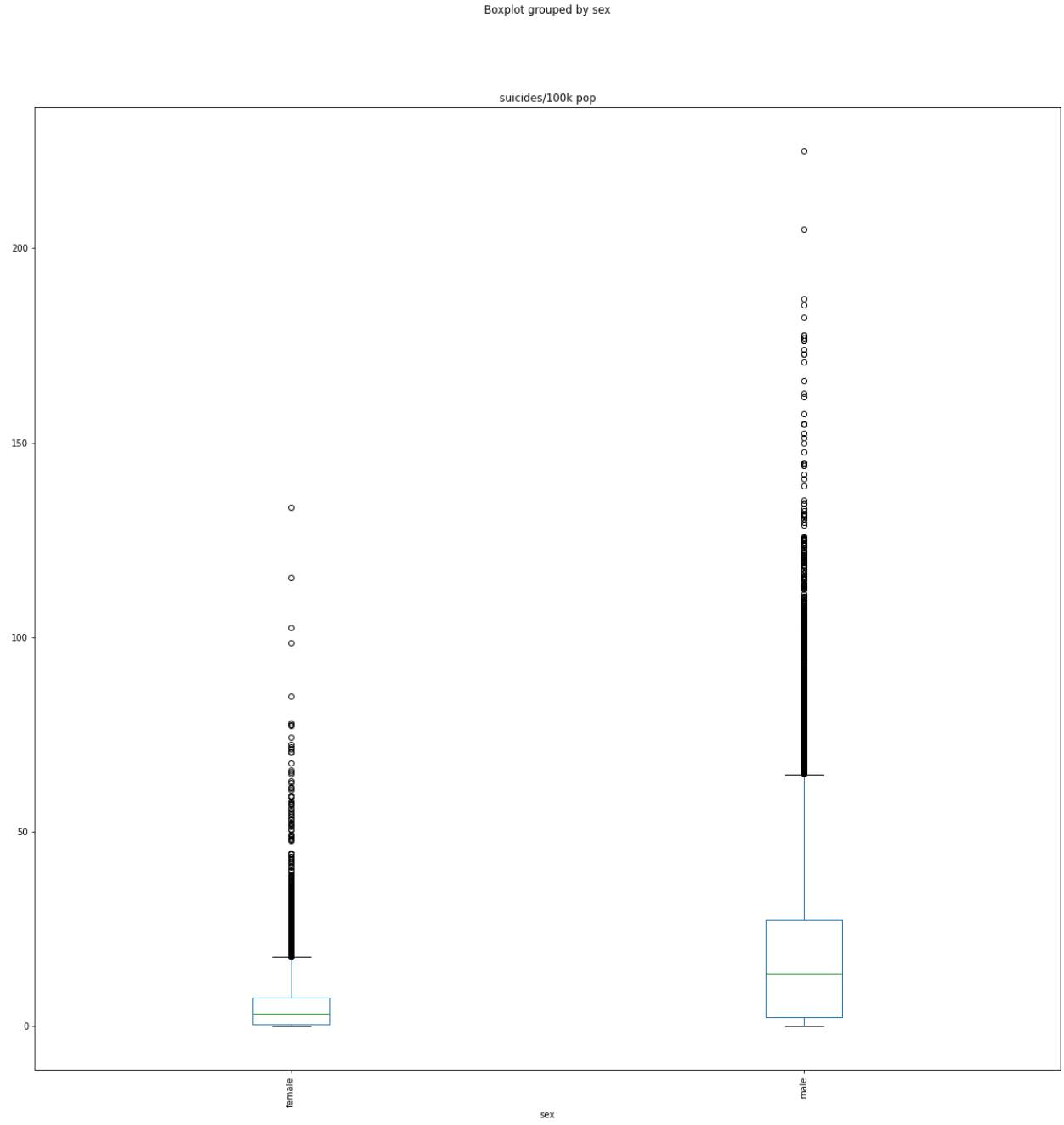
Suicide:

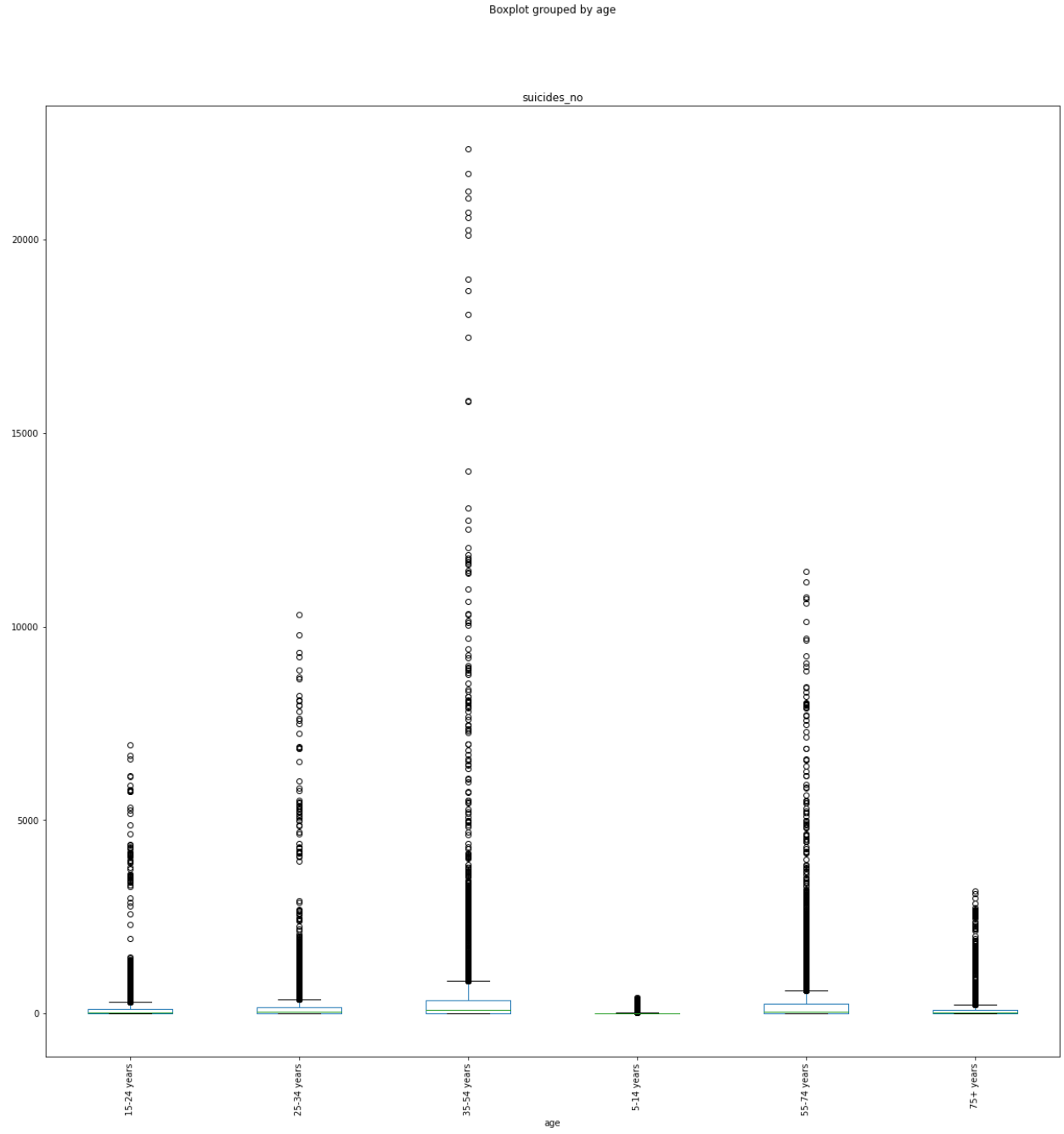
For the boxplots, we can find out that there are more males in suicided people. We cannot exactly tell how many countries have more suicided people / 100k population. Countries like Russian and Ukraine has higher suicide rate.

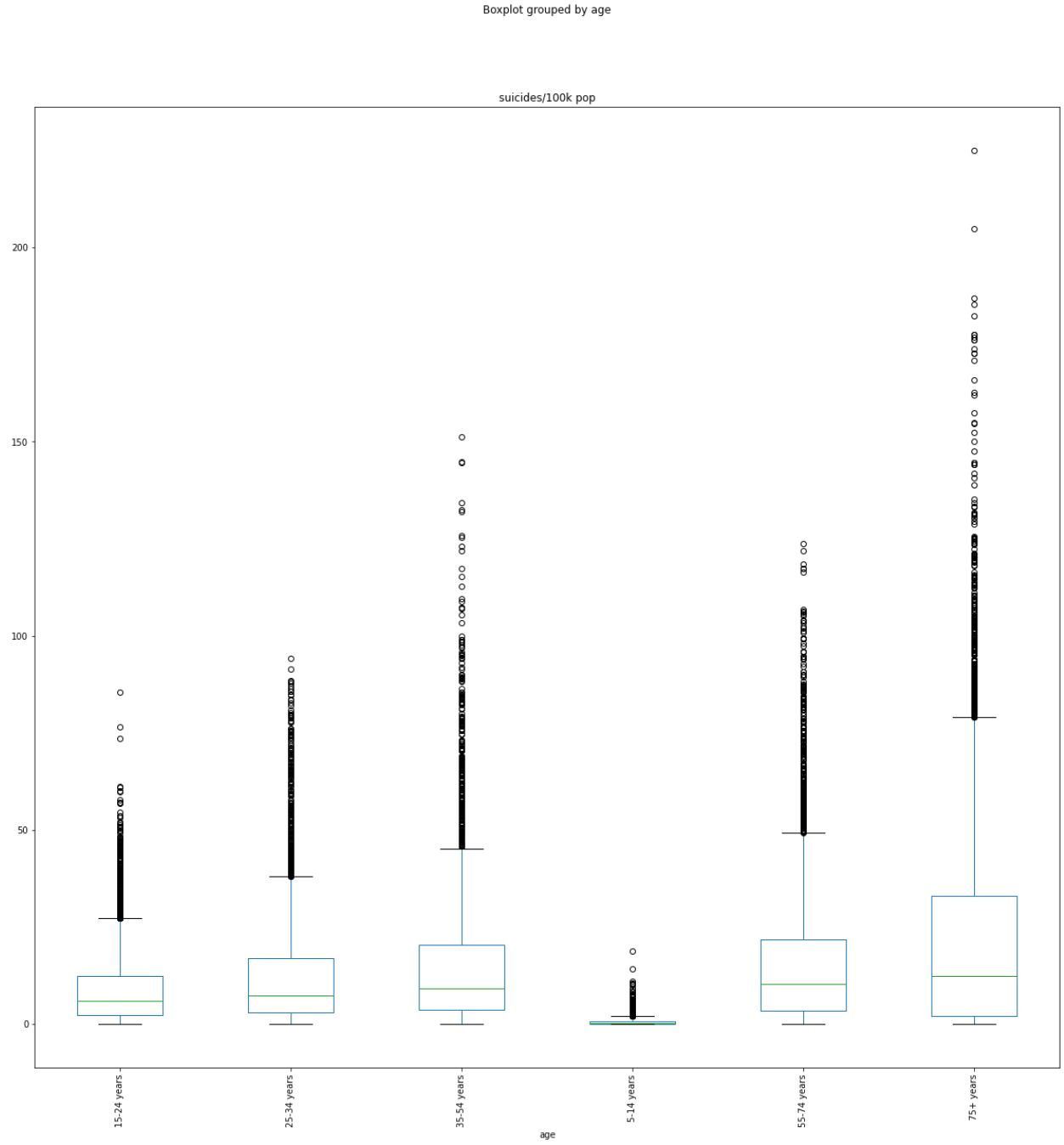


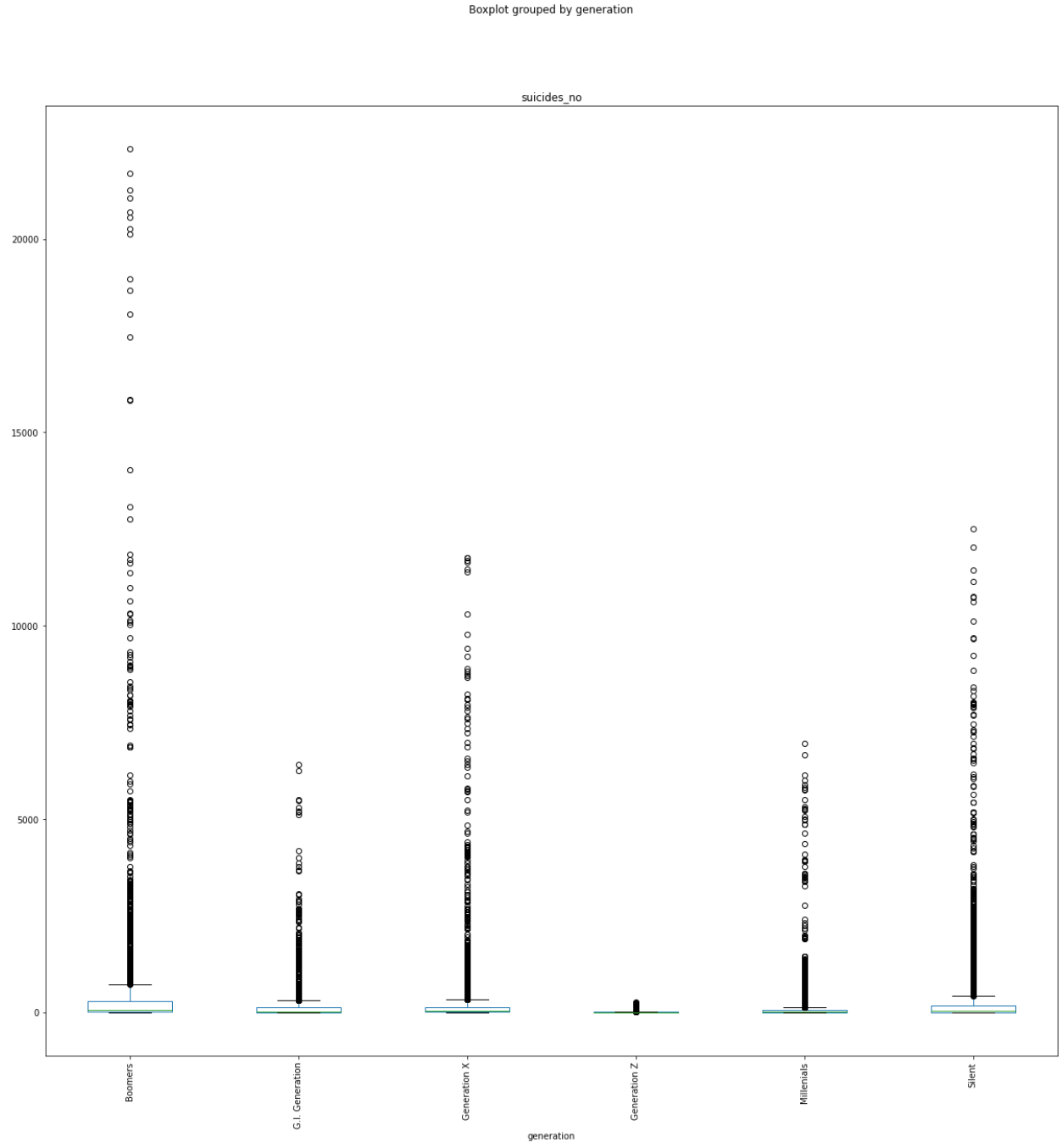

Boxplot grouped by country
suicides/100k pop
0
50
100
150
200
Albania
Antigua and Barbuda
Argentina
Armenia
Aruba
Australia
Austria
Azerbaijan
Bahamas
Bahrain
Barbados
Belarus
Belgium
Belize
Bosnia and Herzegovina
Brazil
Bulgaria
Cabo Verde
Canada
Chile
Colombia
Costa Rica
Croatia
Cuba
Cyprus
Czech Republic
Denmark
Dominica
Ecuador
El Salvador
Estonia
Fiji
Finland
France
Georgia
Germany
Greece
Grenada
Guatemala
Guyana
Hungary
Iceland
Ireland
Israel
Italy
Jamaica
Japan
Kazakhstan
Kiribati
Kuwait
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Macau
Maldives
Malta
Mauritius
Mexico
Mongolia
Montenegro
Netherlands
New Zealand
Nicaragua
Norway
Oman
Panama
Paraguay
Philippines
Poland
Portugal
Puerto Rico
Qatar
Republic of Korea
Romania
Russian Federation
Saint Kitts and Nevis
Saint Lucia
Saint Vincent and Grenadines
San Marino
Serbia
Seychelles
Singapore
Slovakia
Slovenia
South Africa
Spain
Sri Lanka
Suriname
Sweden
Switzerland
Thailand
Trinidad and Tobago
Turkey
Turkmenistan
Ukraine
United Arab Emirates
United Kingdom
United States
Uruguay
Uzbekistan
country

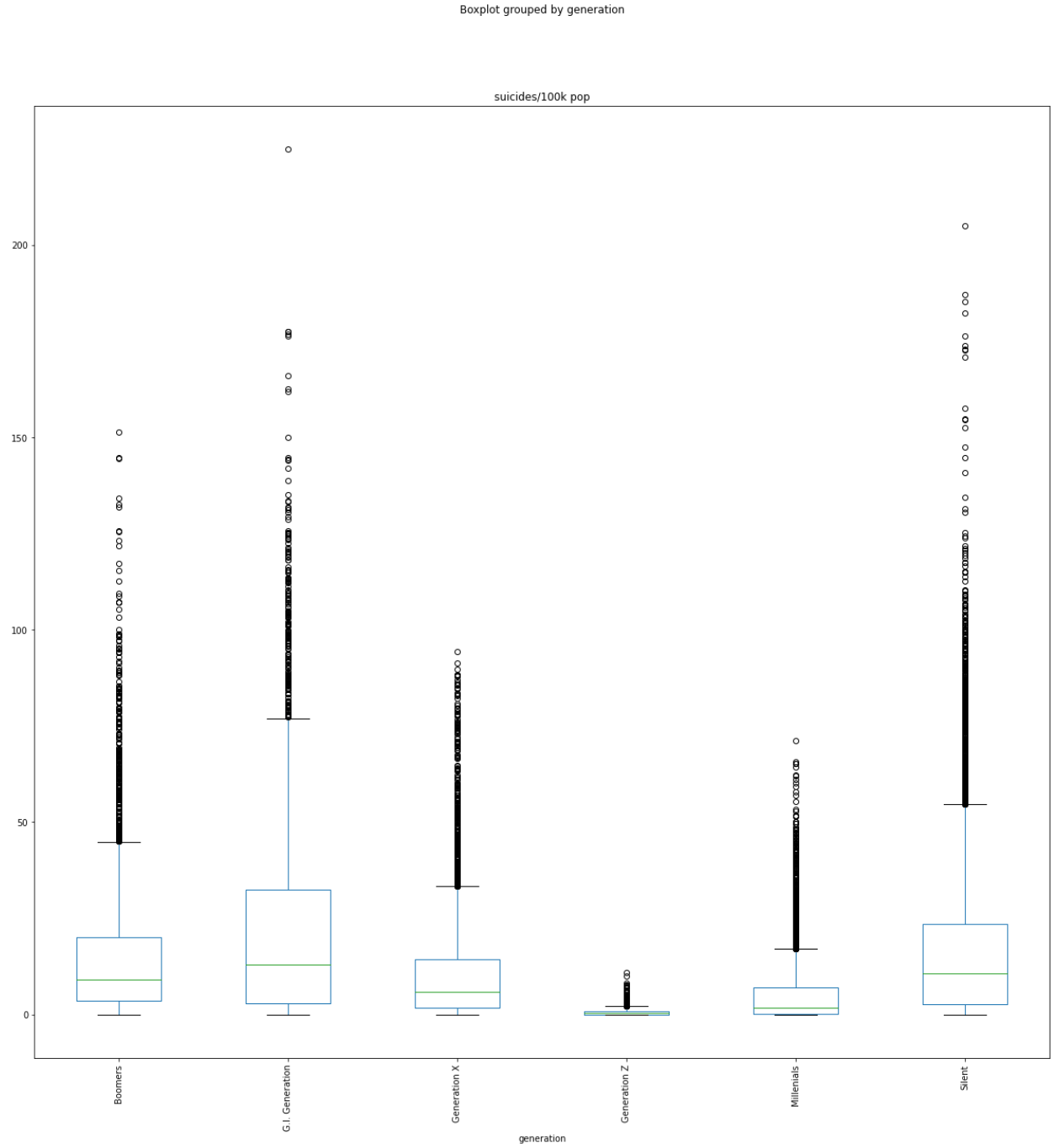

Boxplot grouped by sex
suicides_no
20000
15000
10000
5000
0
female
male
sex


Boxplot grouped by sex
suicides/100k pop
200
150
100
50
0
female
male
sex


Boxplot grouped by age
suicides_no
20000
15000
10000
5000
0
15-24 years
25-34 years
35-54 years
5-14 years
55-74 years
75+ years
age

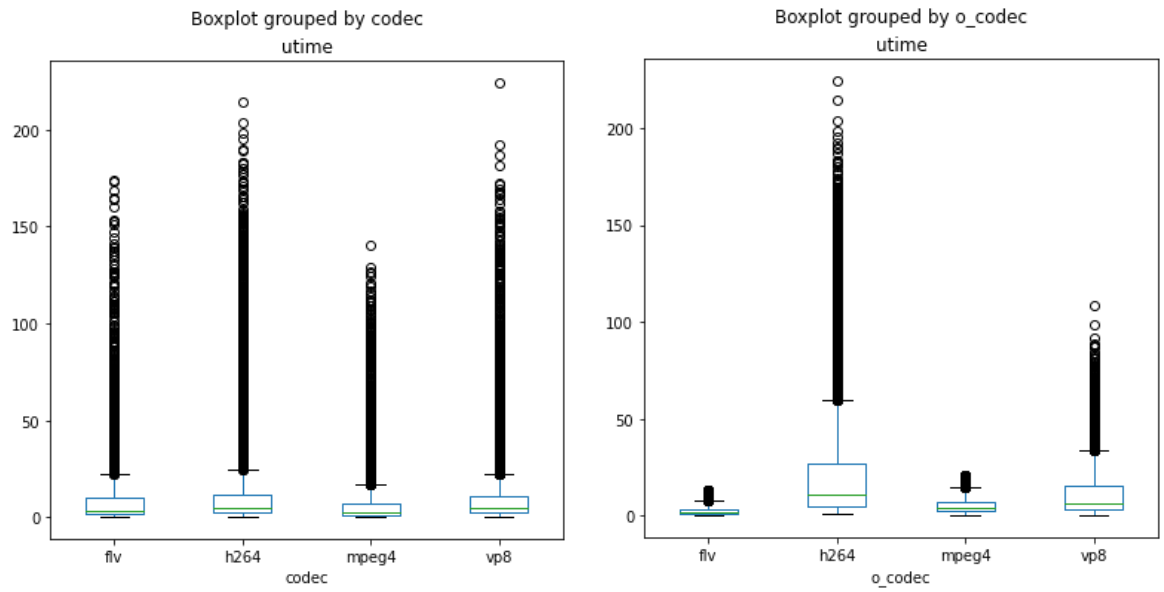

Boxplot grouped by age
suicides/100k pop
200
150
100
50
0
15-24 years
25-34 years
35-54 years
5-14 years
55-74 years
75+ years
age


Boxplot grouped by generation
suicides_no
20000
15000
10000
5000
0
Boomers
G.I. Generation
Generation X
Generation Z
Millenials
Silent
generation


Boxplot grouped by generation
suicides/100k pop
200
150
100
50
0
Boomers
G.I. Generation
Generation X
Generation Z
Millenials
Silent
generation

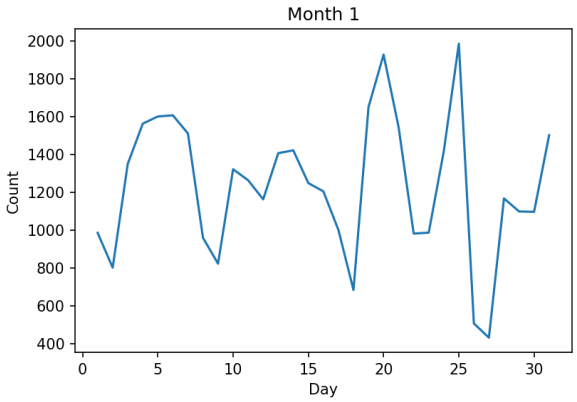
Video:

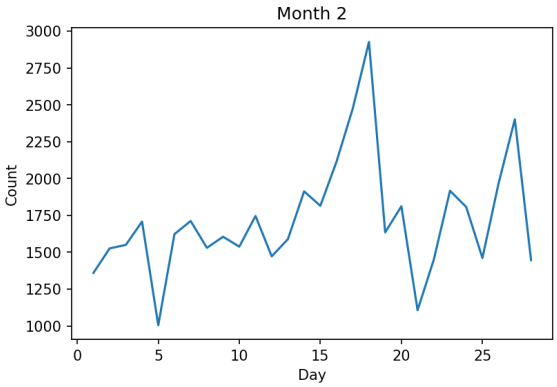
For video data we can see that there are two categorical data, codec and output codec. The utime is much larger if the o_codec is h264. And we can see that there are many outliers in these two categorical data and we need to standardize them to avoid data skewness.

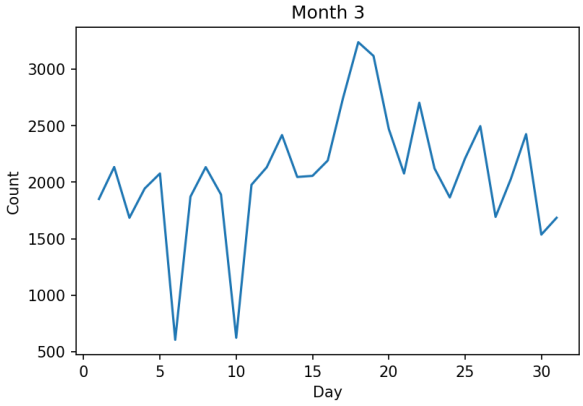


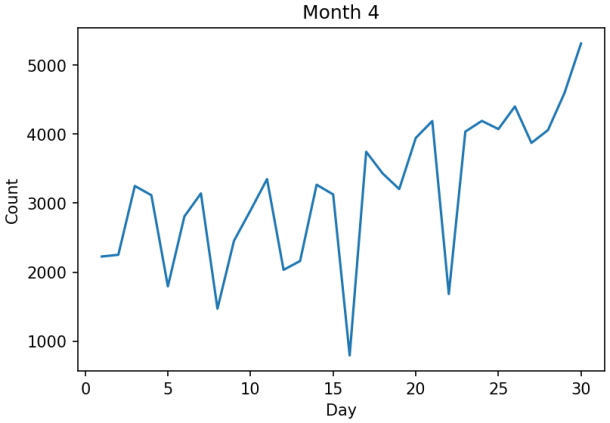
Question 4:

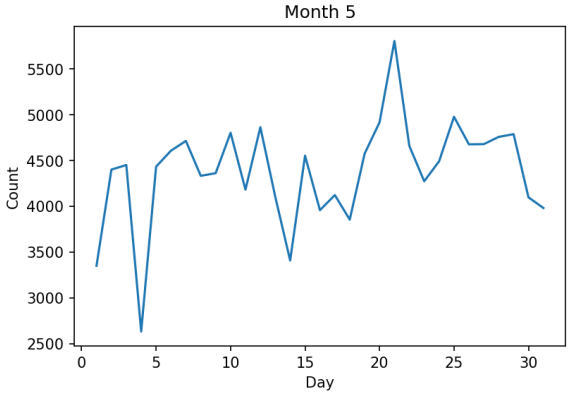
For bike sharing dataset, we have plotted the count number per day for 8 months. And we have observed that there was a repeating pattern that cnt varies a lot from day to day, and the highest count is sometimes on the 15th to 20th, and sometimes around the 30th. The lowest count appears at the beginning of a month and sometimes around the 20th.

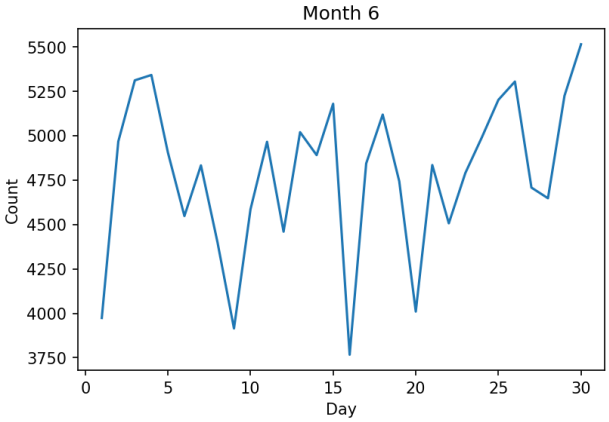

Month 1
Count
Day

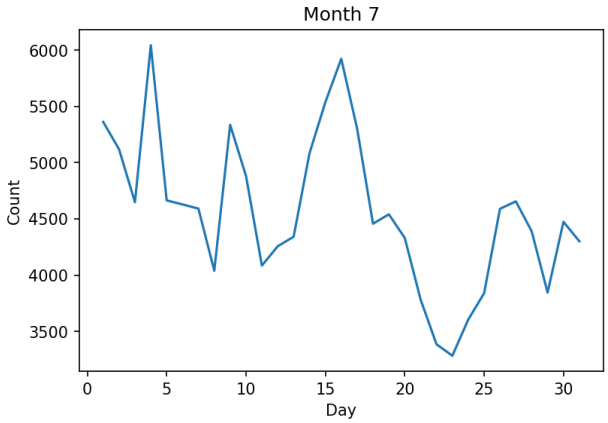

Month 2
Count
Day

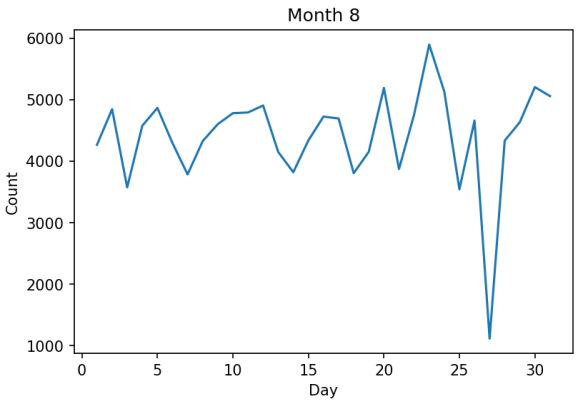

Month 3
Count
Day


Month 4
Count
Day


Month 5
Count
Day

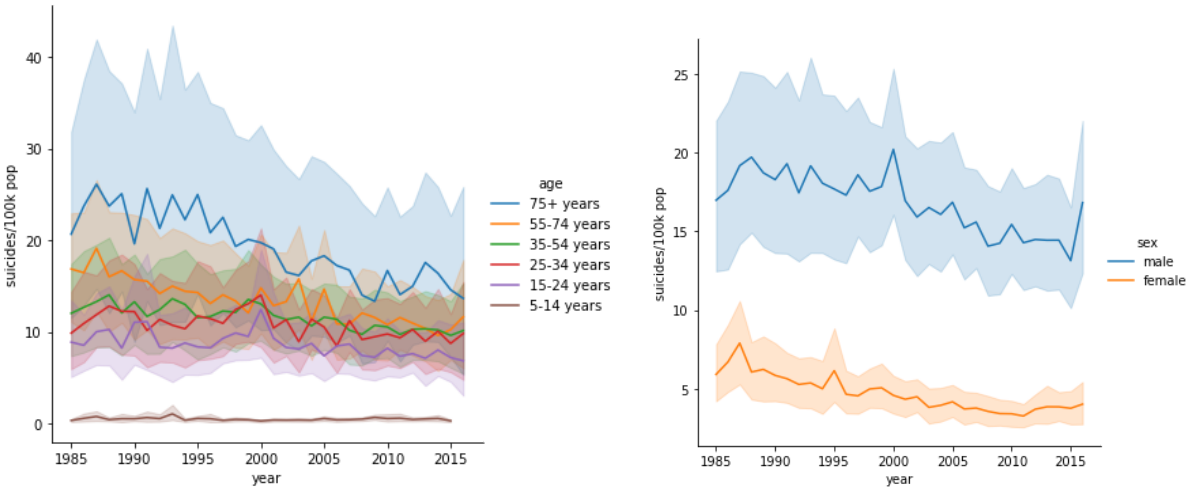

Month 6
Count
Day





Question 5:

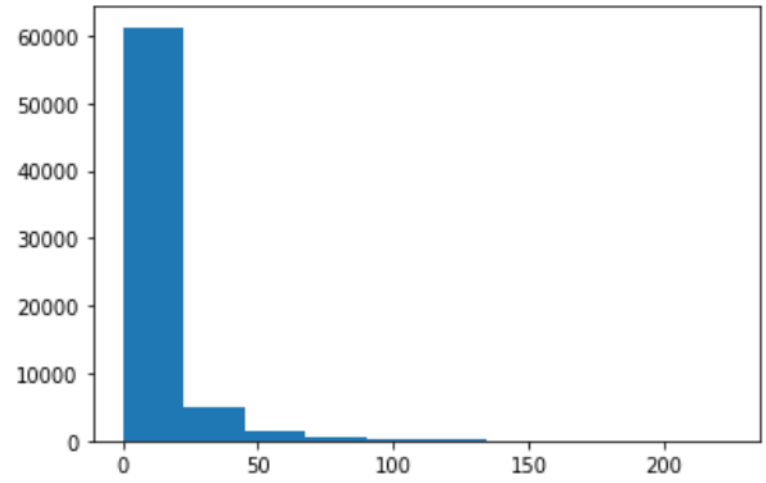
For the suicide rate dataset, we picked the top 10 countries that have the longest timespan of records (in terms of years). We have plotted the suicide rate against time for different age groups and gender. We can see a trend that there are less and less suicided people each year. But 75+ ages and males account for most.



Question 6:

For the video transcoding time dataset, we plotted the distribution of video transcoding times, and we observed that the distribution of is highly skewed, and the transcoding time for most videos is between 0 and 50. And this shape of distribution corresponds to the mean and median transcoding times attached below:

```
The mean of transcoding times is: 9.996354820888516
The median of transoding times is: 4.408
```



Question 7:

A categorical feature is a feature that can take on one of a limited number of possible values. A preprocessing step is to convert categorical variables into numbers and thus prepared for training. One method for numerical encoding of categorical features is to assign a scalar. And another method is to perform one-hot encoding. Specifically, for the suicide rate dataset, the number of unique countries for the variable "country" is pretty high. So we grouped these countries into same continent countries such as Europe, North America, South America, Middle East and Asia, and proceeded with one-hot encoding. For one-hot encoding, we can apply it to cases with no significance difference, which means that every element in this ease has the same significance, and therefore discarding the ordering information. However, for scalar encoding, there exists an order or significance difference among the elements, and we retain this ordering information. Therefore, we applied pd.get_dummies to perform one-hot encoding for all three datasets because those columns share the same importance. For suicide dataset, we convert all countries into their continents they located at which help reduce too much encodings to avoid

multicolinearity with other one-hot encoding features. For scalar encoding, we need to assume there is a ranking relationship between certain categorical data.

Question 8:

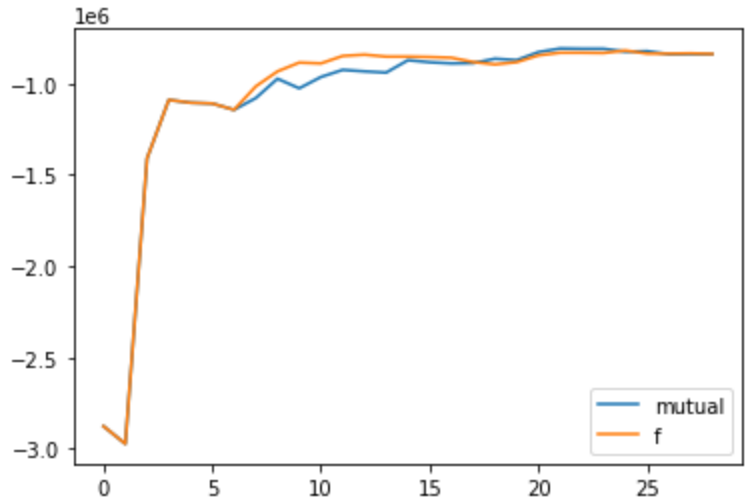
Standardization of datasets is a common requirement for many machine learning estimators; they might behave badly if the individual features do not more-or-less look like standard normally distributed data: Gaussian with zero mean and unit variance. If a feature has a variance that is orders of magnitude larger than others, it might dominate the objective function and make the estimator unable to learn from other features correctly as expected. Therefore,our team standardized feature columns and prepared them for training by using StandardScalar().fit_transform for the training data of all 3 datasets. After that, each column in the dataset had zero mean and unit variance, just like a standard normal distribution.

Question 9:

In this step, we have used 2 functions to select the most important features. The first one is called sklearn.feature selection.mutual info regression. It returns estimated mutual information between each feature and the label. Mutual information (MI) between two random variables is a non-negative value which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency.

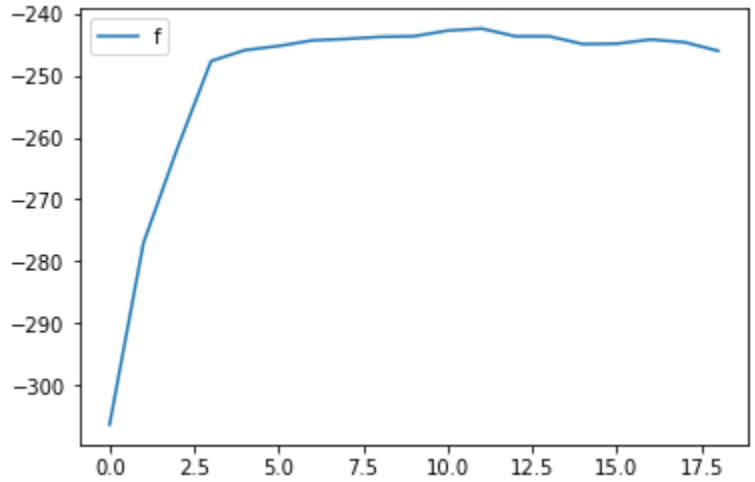
Another function is called sklearn.feature selection.f regression. It provides F scores, which is a way of comparing the significance of the improvement of a model, with respect to the addition of new variables. This step would help avoid overfitting and greatly improve the performance of the models in terms of test RMSE. The results are attached below:

Bike:



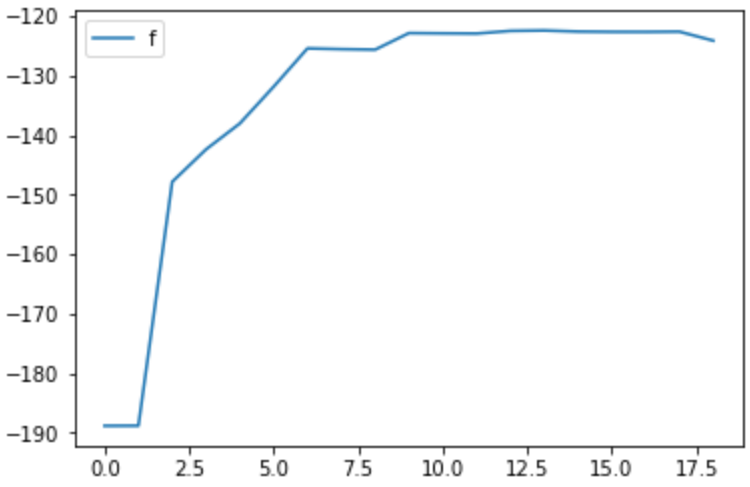
And we have found the best mu score to be 20, and the best f score to be 24.

Suicide:



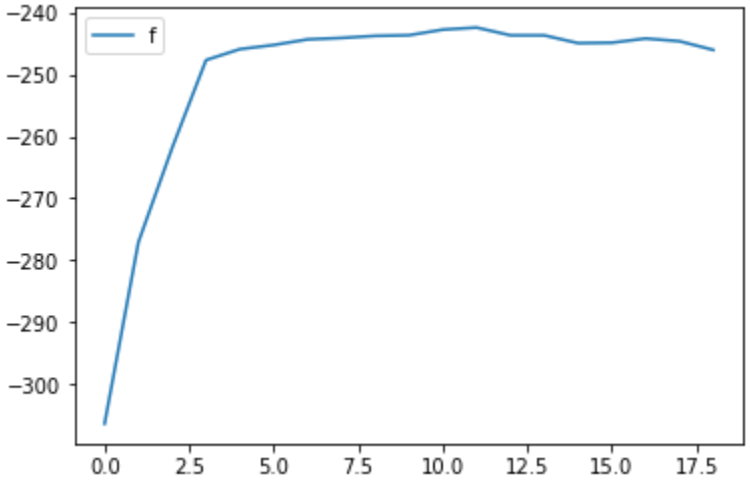
We have found the best f score to be 11.

Video:



We have found the best f score to be 13.

Suicide:



We have found the best f score to be 11.

Question 10

Compared to linear regression, Lasso regression requires L1 norm and Ridge regression requires Ridge norm. The difference is that L1 regularization uses only a part of the model while L2 utilizes all the features when learning. This difference makes L1 (Lasso) better for screening and L2 (Ridge) better for shrinking.

Question 11

	mean_test_score	mean_train_score	param_model	Standardize
0	-11.037695	-11.024954	Lasso(alpha=0.1, copy_X=True, fit_intercept=Tr...	1
1	-11.052017	-11.000246	Lasso(alpha=0.1, copy_X=True, fit_intercept=Tr...	1
2	-11.054490	-10.997239	Lasso(alpha=0.1, copy_X=True, fit_intercept=Tr...	1
3	-11.055412	-10.997215	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
4	-11.055838	-10.997195	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
5	-11.055909	-10.997201	Lasso(alpha=0.001, copy_X=True, fit_intercept=...	0
6	-11.056076	-10.997175	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
7	-11.056641	-10.997452	Lasso(alpha=0.001, copy_X=True, fit_intercept=...	0
8	-11.057089	-10.997137	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
9	-11.058241	-10.997130	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
10	-11.058560	-10.997130	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	0

Linear Regression Model for Bike Sharing Dataset

	mean_test_score	mean_train_score	param_model	Standardize
0	-11.037695	-11.024954	Lasso(alpha=0.1, copy_X=True, fit_intercept=Tr...	1
1	-11.052017	-11.000246	Lasso(alpha=0.1, copy_X=True, fit_intercept=Tr...	1
2	-11.054490	-10.997239	Lasso(alpha=0.1, copy_X=True, fit_intercept=Tr...	1
3	-11.055412	-10.997215	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
4	-11.055838	-10.997195	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
5	-11.055909	-10.997201	Lasso(alpha=0.001, copy_X=True, fit_intercept=...	0
6	-11.056076	-10.997175	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
7	-11.056641	-10.997452	Lasso(alpha=0.001, copy_X=True, fit_intercept=...	0
8	-11.057089	-10.997137	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
9	-11.058241	-10.997130	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
10	-11.058560	-10.997130	Ridge(alpha=1.0, copy_X=True, fit_intercept=Tr...	0

Linear Regression Model for Video Dataset

	mean_test_score	mean_train_score	param_model	Standardize
0	-15.297404	-15.393794	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
1	-15.303900	-15.392730	Lasso(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
2	-15.304535	-15.392536	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
3	-15.305343	-15.392524	Lasso(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
4	-15.305420	-15.392522	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
5	-15.305510	-15.392521	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
6	-15.305519	-15.392521	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
7	-15.305520	-15.392521	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
8	-15.308852	-15.413462	Lasso(alpha=1.0, copy_X=True, fit_intercept=Tr...	1
9	-15.316081	-15.471623	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	1
10	-15.324642	-15.424555	Ridge(alpha=100.0, copy_X=True, fit_intercept=...	0

Linear Regression Model for Suicide Dataset

In this problem, we used grid search to optimize the regulation and applied cross validation to find the highest root mean square errors. All three datasets are applied with this technique and the result table is attached.

Question 12

Feature scaling will not play a role without regularization. The standardization effect on linear regression can prove that only one feature change will not affect the whole result. If regularization is performed, the lambda coefficient will make a difference. This can be proved by the Lasso and Ridge regularization.

Question 13

P-values are the probability of feature coefficient to be zero. So if p-value is small then the feature is making a big difference in the model. We used `statsmodels.api` to measure p-values and the results below is the p-values for bike share and video dataset:

```
const            1.291438e-11
yr              2.294966e-154
holiday          2.575233e-01
workingday       2.896613e-11
temp             4.152649e-02
atemp            2.222607e-01
hum              2.013660e-07
windspeed        2.091887e-11
weekday_0        6.595401e-02
weekday_1        2.725026e-01
weekday_2        7.741005e-01
weekday_3        2.039693e-01
weekday_4        1.748603e-01
weekday_5        3.959554e-02
weekday_6        1.483723e-13
season_1         2.656985e-06
season_2         9.671326e-03
season_3         4.188569e-02
season_4         6.407070e-15
mnth_1           1.286778e-01
mnth_2           4.217047e-01
mnth_3           2.533052e-02
mnth_4           2.151364e-01
mnth_5           3.267647e-03
mnth_6           1.002822e-01
mnth_7           2.031799e-01
```

```
mnth_8           2.932908e-01
mnth_9           8.327337e-08
mnth_10          9.782034e-02
mnth_11          2.110197e-02
mnth_12          1.352225e-02
weathersit_1     5.481062e-52
weathersit_2     6.510439e-18
weathersit_3     8.569032e-08
dtype: float64
```

P-values for bike dataset

```
const           5.235365e-190
duration         4.709001e-02
width            5.568883e-08
height           9.271988e-09
bitrate         1.456482e-126
framerate        3.292809e-10
i                8.259887e-08
p                1.094799e-04
b                7.069405e-06
frames           2.192189e-04
i_size           4.681481e-01
p_size           3.821872e-01
size             3.846849e-01
o_bitrate        0.000000e+00
o_framerate      0.000000e+00
o_width         5.218535e-130
o_height         1.108057e-03
codec_flv        8.780535e-16
codec_h264       1.467975e-12
codec_mpeg4     3.905647e-129
codec_vp8        5.592198e-30
o_codec_flv      0.000000e+00
o_codec_h264     0.000000e+00
o_codec_mpeg4    0.000000e+00
o_codec_vp8      7.954758e-10
dtype: float64
```

P-values for video dataset

```
year                          8.790827e-04
population                    2.072198e-03
 gdp_for_year ($)             7.528723e-11
gdp_per_capita ($)            2.090879e-47
continent_Asia                4.696579e-11
continent_Europe              8.209215e-88
continent_North America       4.747729e-01
continent_Oceania             7.666389e-12
continent_South America       1.048222e-06
sex_male                      0.000000e+00
age_25-34 years               1.913446e-19
age_35-54 years               1.531597e-27
age_5-14 years               5.950804e-112
age_55-74 years               1.520674e-21
age_75+ years                 2.233077e-54
generation_G.I. Generation    7.043793e-01
generation_Generation X       3.364701e-01
generation_Generation Z       1.076723e-02
generation_Millenials         2.794787e-01
generation_Silent             2.439058e-02
```

P values for suicide dataset

Question 14

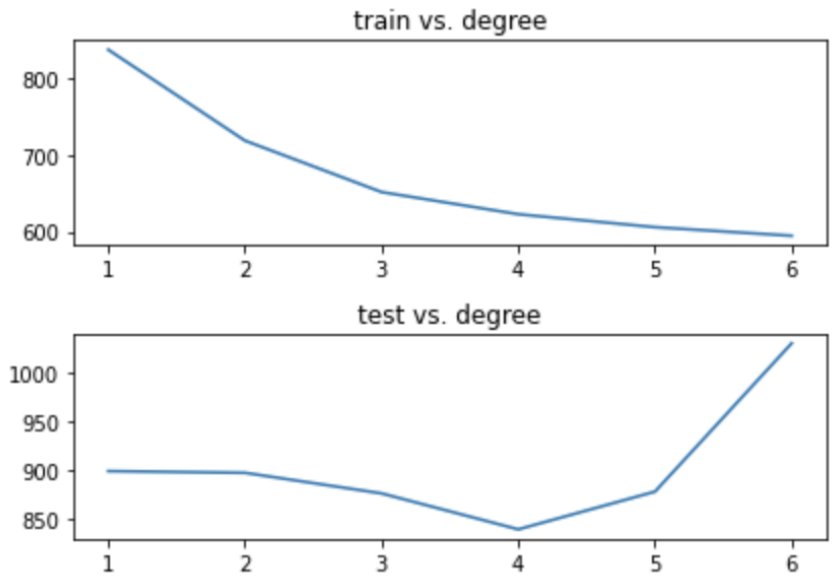
Salient features have larger absolute parameter values. For the bike dataset, we found out that temp and atemp have the highest absolute values with the target variable. It means that more people are more willing to rent a bike in nice weather days.

For the video dataset, we found out o_width and o_height are salient features. It means that the video size is a big factor when we consider the transcoding time.

For the suicide dataset, we found out population and gdp_for_year are salient features, which is also reasonable. Rich countries usually are capitalist society, meaning that bottom level people are exploited by top level people, which might be an important factor.

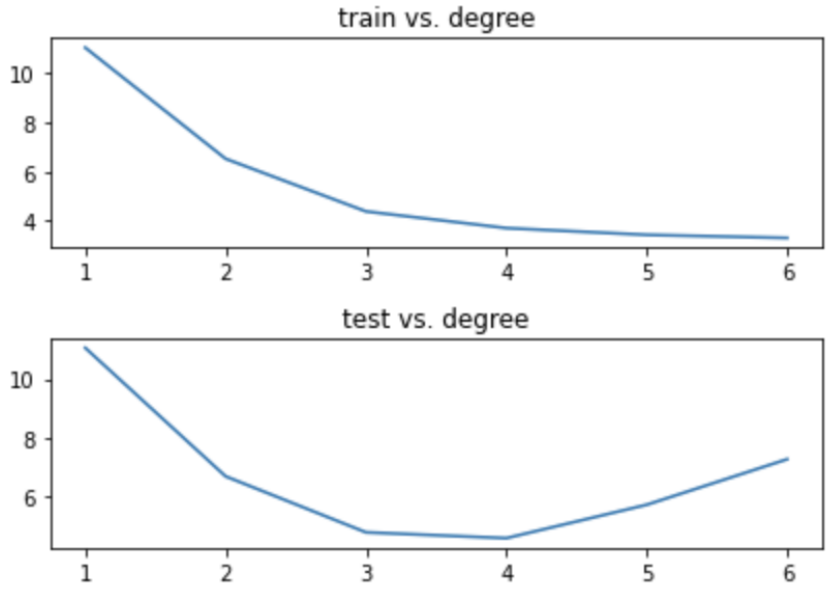
Question 15

	param_poly__degree	mean_test_score	mean_train_score
0	1	-898.352426	-836.936914
1	2	-896.613639	-718.505359
2	3	-875.469510	-651.545718
3	4	-838.575823	-622.665266
4	5	-877.298241	-605.808878
5	6	-1029.552275	-594.662188



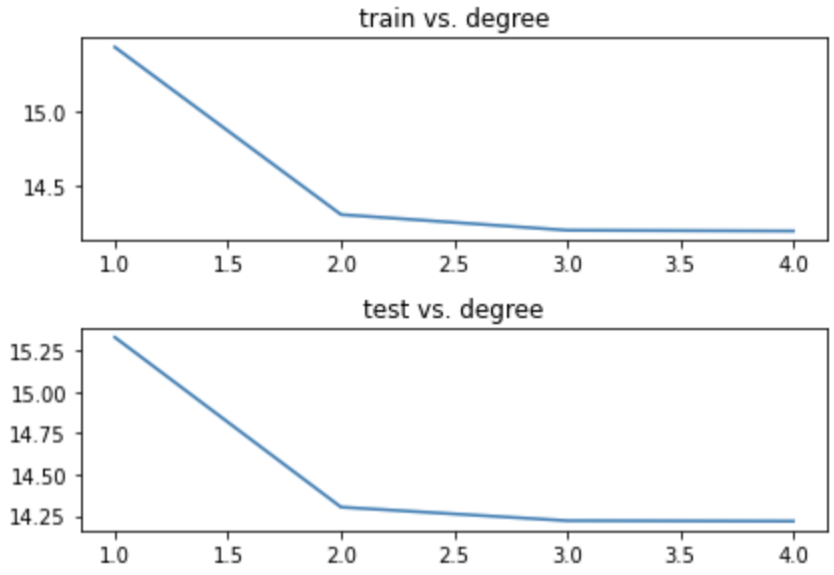
Bike Dataset Standardization Results

	param_poly__degree	mean_test_score	mean_train_score
0	1	-11.059453	-11.038711
1	2	-6.672095	-6.524858
2	3	-4.759948	-4.386742
3	4	-4.561423	-3.708211
4	5	-5.702094	-3.434984
5	6	-7.256978	-3.309889



Video Dataset Standardization Results

	param_poly__degree	mean_test_score	mean_train_score
0	1	-15.330860	-15.423485
1	2	-14.303820	-14.310242
2	3	-14.221883	-14.205939
3	4	-14.218914	-14.200961



Suicide Dataset Standardization Results

The optimal degree of polynomial for bike and video is 4 since after that the root mean square error starts to increase. The optimal degree of polynomial for suicide is 3 since even though the error keeps decreasing after that but the gain is marginal so we do not need to choose a higher degree.

Question 16

Since all the related factors are in the numerator and denominator, the transcoding accuracy is directly proportional to these factors. We get a new feature as o_height * o_width / o_bitrate
In the experiment we did, we successfully decreased the root mean square error to 11.056075997583076.

Question 17

From the RMSE we got on both datasets we can see that the perceptron neural network is better because it can interpret non-linear data while the linear regression, by its name, can only deal with linear features and targets. Also, with back propagation, the SGD algorithm can help reach global minimum and minimize the loss function.

Question 18

	param_alpha	param_activation	param_hidden_layer_sizes	mean_test_score	mean_train_score
4	0.001	relu	(32, 32, 32)	-818.914623	-535.584457
5	0.001	relu	(64, 64, 64)	-819.798967	-477.207525
6	0.01	relu	(32, 32, 32)	-839.783469	-518.656826
7	0.01	relu	(64, 64, 64)	-851.025704	-452.807060
2	0.01	identity	(32, 32, 32)	-868.211400	-760.045527

Bike Dataset Grid Search

	param_alpha	param_activation	param_hidden_layer_sizes	mean_test_score	mean_train_score
4	0.001	relu	(32, 32, 32)	-4.704740	-2.625139
6	0.01	relu	(32, 32, 32)	-4.763165	-2.617754
5	0.001	relu	(64, 64, 64)	-5.142316	-1.875029
7	0.01	relu	(64, 64, 64)	-5.642840	-1.798004
1	0.001	identity	(64, 64, 64)	-11.009360	-11.029723

Video Dataset Grid Search

	param_alpha	param_activation	param_hidden_layer_sizes	mean_test_score	mean_train_score
3	0.01	relu	(32, 32, 32)	-14.129354	-14.217982
2	0.1	relu	(32, 32, 32)	-14.150664	-14.215628
1	0.01	identity	(32, 32, 32)	-15.309918	-15.441639
0	0.1	identity	(32, 32, 32)	-15.334521	-15.443702

Suicide Dataset Grid Search

Question 19

Theoretically, we were performing regression here so no activation function is needed. But we noticed that with relu as an activation function our model got a better test score. It might because Relu can handle nonlinear features and avoid vanishing gradient issues of the neural network.

Question 20

Firstly, if the neural network is too deep, we will come across overfitting problems and it is hard for neural networks to learn identity mapping. Secondly, it will be very difficult to train the neural network because there are much more parameters which is very time consuming.

Question21

For the bike dataset, the best mean test score and hyperparameters we got are:

`n_estimators=110,max_features=17,max_depth=20`

RMSE= 820.109910

For the video dataset, the best mean test score and hyperparameters we got are:

`n_estimators=190,max_features=5,max_depth=21`

RMSE = 3.998347

For the bike dataset, the best mean test score and hyperparameters we got are:

`n_estimators=150,max_features=3,max_depth=13`

RMSE = 14.208148

Max_features is the size of the random subsets of features to consider when splitting a node. Number of trees is the bagging method of random forest. More trees also mean more computational cost and after a certain number of trees, the improvement is negligible. Max_depth can help avoid any specific tree select all features which may overfit the data.

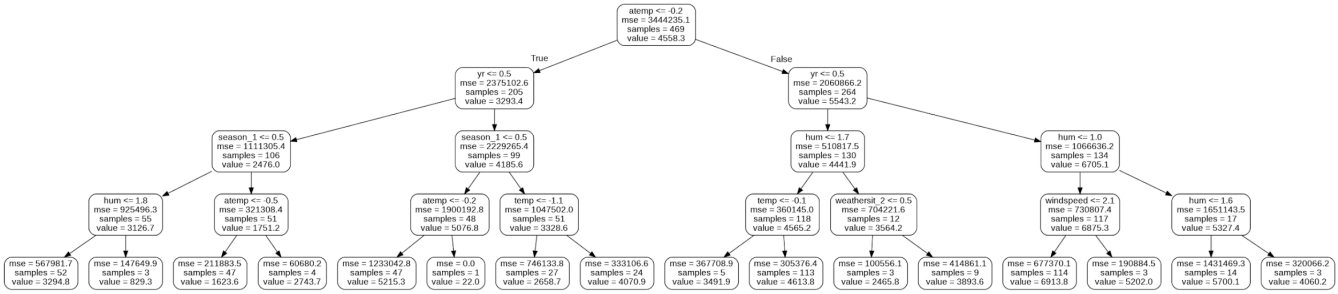
Question 22

From our testing, we can see that Random Forest Regressor gives the best RMSE over all techniques. Random Forest uses bootstrap and bagging methods that only use a subset of dataset with replacement and is more robust to missing data and avoid overfitting because it reduces correlation between trees.

Question 23

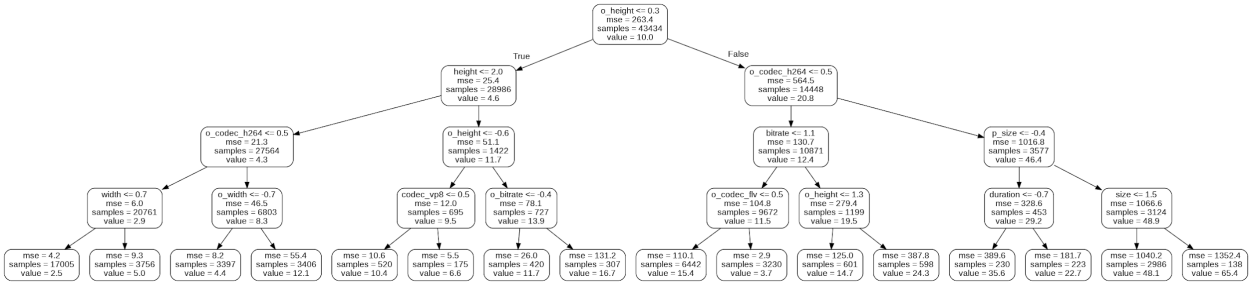
Random Forest for Bike Dataset

We can see that the root node is atemp, and it is the most important feature in this tree. Also, in 3.2.1 we conclude that atemp is one of the most important feature of the bike dataset.



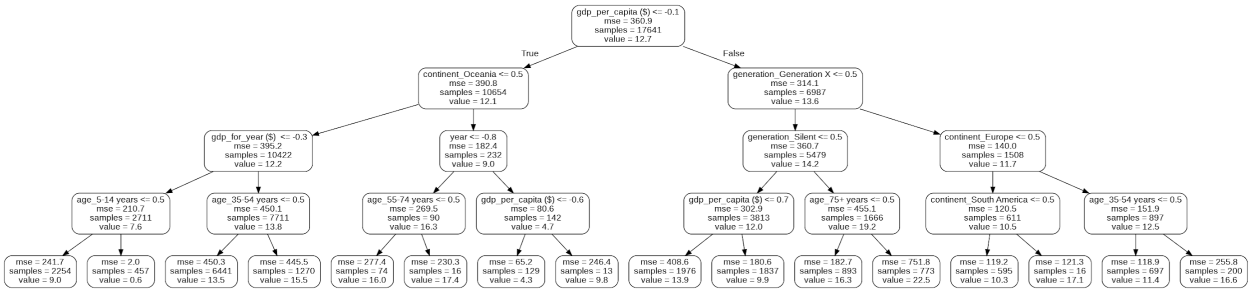
Random Forest for Video Dataset

We can see that the root node is o_height, and it is the most important feature in this tree. Also, in 3.2.1 we conclude that o_height is one of the most important features of the video dataset.



Random Forest for Suicide Dataset

We can see that the root node is gdp_per_capita, and it is the most important feature in this tree. Also, in 3.2.1 we conclude that gdp is one of the most important features of the suicide dataset.



Question 24

We decided to work on the bike dataset because it is smaller than the other two datasets which help reduce the training time. For the LightGBM parameters we are searching for are

```
num_leaves":(10,50),
n_estimators":(10,120),
max_depth":(1,30)
```

For CatBoost we are searching for

```
learning_rate':(0.01, 1.0, 'log-uniform'),
model__depth': (1, 4),
model__l2_leaf_reg': (2, 30),
```

And we decided to use BayesSearchCV for finding the best hyperparameter combinations which can be referred to in the next question.

Question 25

For LightGBM: The best combination we found is with 11 leaves, 69 estimators and max depth at 15.

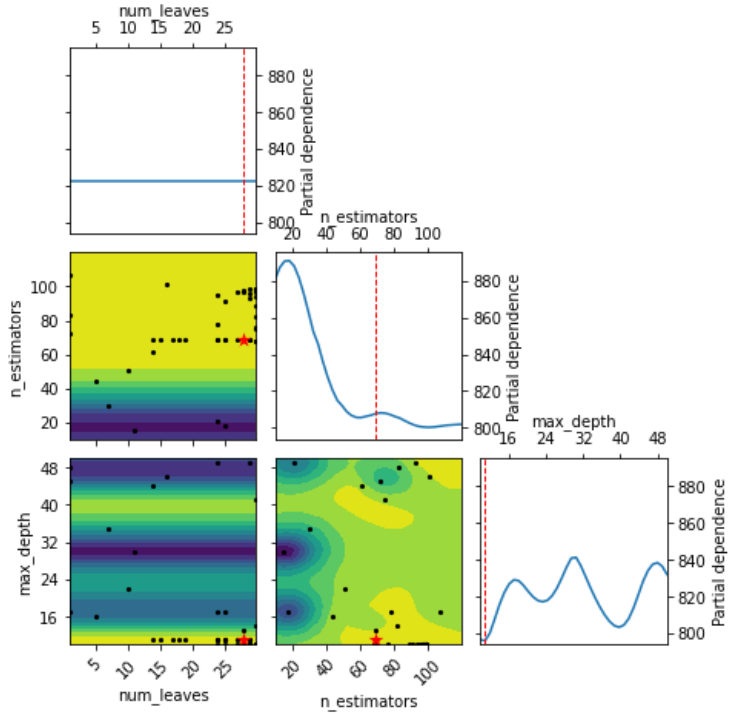
	mean_test_score	param_model__num_leaves	param_model__n_estimators	param_model__max_depth
49	-779.789132	11	69	15
42	-779.789132	11	69	14
32	-779.789132	11	69	29
16	-779.789132	11	69	28
35	-779.789132	11	69	28

For CatBoost: The best combination we found is with 0.018377 learning rate, model depth at 2 and coefficient at the L2 regularization term of the cost function with a value of 2.

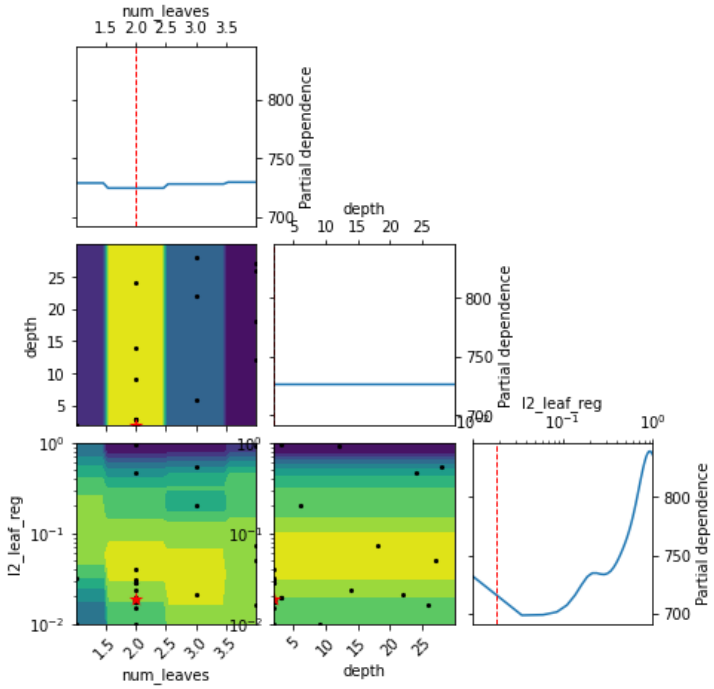
	mean_test_score	param_model__learning_rate	param_model__depth	param_model__l2_leaf_reg
39	-677.469871	0.018377	2	2
46	-678.125585	0.018454	2	2
42	-678.357303	0.018345	2	2
26	-678.359336	0.018063	2	2
30	-678.373736	0.017782	2	2

Question 26

Plot for LightGBM:



Plot for CatBoost:



From the mean_test RMSE in question 25 we can see that CatBoost helps reduce RMSE a lot for nearly 150 and LightGBM only reduces RMSE around 50. Also, from the two partial dependence plots we can see that for LightGBM the n_estimators affect a lot, and the coefficient at the L2 regularization term of the cost function in CatBoost affects a lot.

Question 27:

We already put calculated RMSE values in previous tables. The RMSE between training and validation data is definitely different. First, we can see that the training error keeps decreasing because the model was trained on training data, but the RMSE for validation decreased first and increased after an optimal value. That is overfitting. Because our model overfit the training data, and is not general to predict validation or test dataset.

Question 28:

For bike dataset, R^2 is `0.983693326828568`, and oob score is `0.8810767540319827`.
For video dataset, R^2 is 0.9988200855497541, and oob score is 0.9922150918129832.
For suicide dataset, R^2 is 0.8365324477769035, and oob score is 0.737444527765355

OOB score is Out-of-Bag score. For random forest, we are using bootstrap techniques thus we are not using the whole dataset. And the OOB score is computing the score of the unused dataset, which is similar to cross validation.

R2 score is the coefficient of determination. It provides a measure of how well a model fits samples. It shows if the observed variation can be explained by the model.