CS 5821
January 24, 2017
Mariia Kravtsova
Assignment 1

1.

- **a.** Flexible model will perform better because the chances of overfitting is lower with an extremely large sample size, so the fit will be closer.
- **b.** Flexible model is likely to overfit the data, so the flexible model is far better. Also, would be better for interpretation of results.
- **c.** To find highly non-linear response one would need to use a flexible model. So a flexible model would be better in this case.
- **d.** Inflexible model would ignore more noise that is provided by high variance. Hence, it is better in this case.

2.

- **a.** The problem is a regression, and the result is of inference type.
  n: 500 (firms); p: 3 (profit, number of employees, industry)
- **b.** This example is a classification problem, in which our goal is to make predictions based on classes of the products.
  n: 20 products; p: 13 (price charged, marketing budget, competition price, and then ten more variable)
- **c.** The nature of the problem is regression, with interest in prediction.
  n: 52 (weekly data for 2012); p: 3 (% change in dollar, % change in British market, % change in German market)

4.

- **a.** 1. On a rough or smooth terrain, make prediction if the car will drive fast or slow? (this applies to training self driving cars). This is a prediction type of problem.

2. Given person’s spending habits by categories (gas, rent, food, etc) and income, is a person's credit score good, bad, or average? Prediction type of problem since we are interested in making conclusions about what affects spending and income have on credit score. This is useful for banks giving loans to a person.

3. Before considering patients for drug trials, we could make predictions if they’ll have positive or negative reaction given their blood type, blood pressure, physical activity, etc. In this case the goal is to make predictions if the drug is successful or not.

b. 1. Based on the average temperatures from past years on this day, what will the weather be today? Response would be an estimated temperature, based on the predictors which are temperatures from previous years. Result type will be prediction.

2. What impact graduate’s GPA has on their salary at their first job after graduation? Response would be different salary rates, based on predictors that would be GPA scores, this is an inference type of problem.

3. Given the amount of carbohydrates consumed daily over the past year by people in US, what is their predicted weight? Weight would be response, predictors would be carbohydrates, and the problem goal is to make a prediction.

c. 1. Based on user behaviours on Netflix, how can we cluster the users and make suggestions of categories to watch next? Where the problem is of inference type, response is categories (Romance, Comedy, Drama, etc), and the predictors would be previous behaviours such as liking particular movies, or watching particular trailers.

2. Given Hubble data from the deep field, categorize it into homogeneous groups for further study. Response - categorized objects (black holes, stars, planets, etc.), and the predictors would be light data of those object in different wavelengths. Results of prediction type.

3. The similarity of genetic data is used in clustering to infer population structures. Response would be population structures, the problem is of inference type, and the predictors are genetic data.

9.

a. Qualitative: name, origin; Quantitative: mpg, cylinders, displacement, horsepower, weight, acceleration, year.

b.

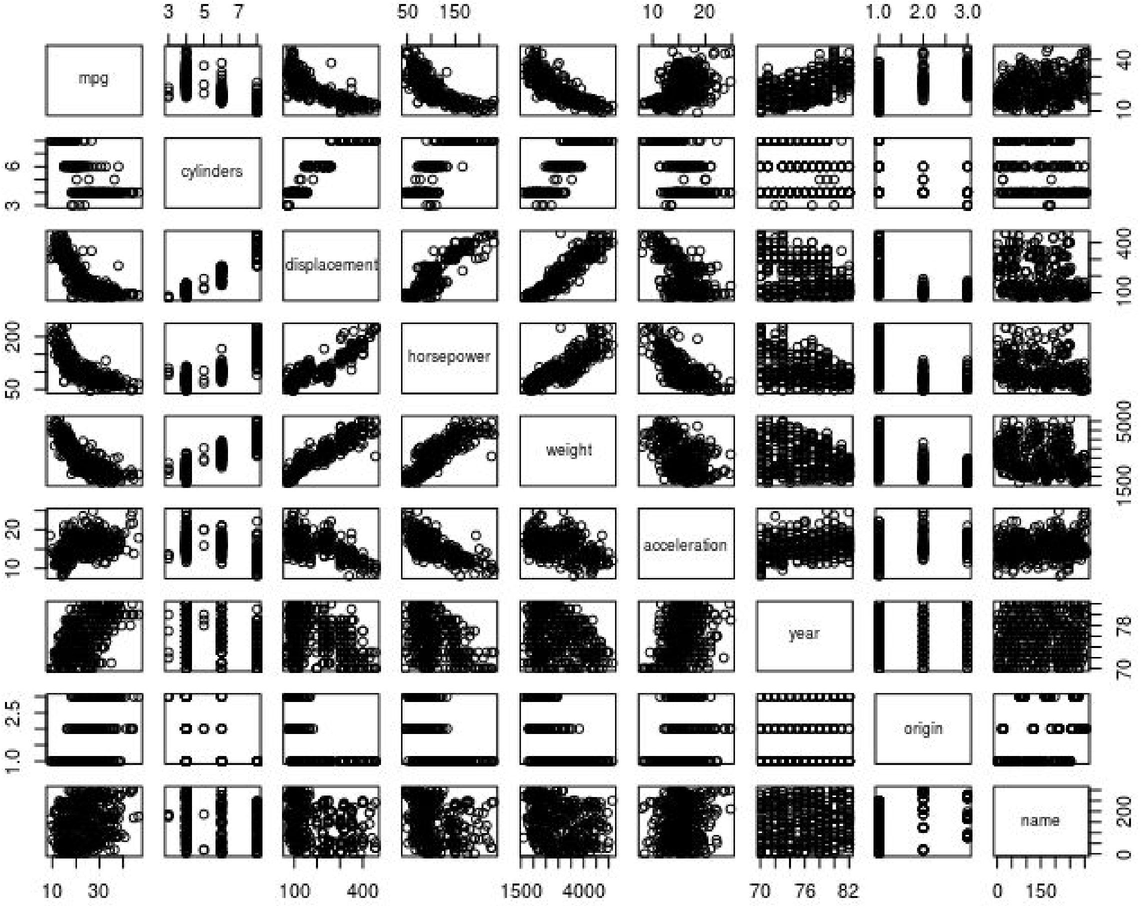
```
> range(mpg)
[1]  9.0 46.6
> range(cylinders)
[1] 3 8
> range(displacement)
[1]  68 455
> range(horsepower)
[1]  46 230
> range(weight)
[1] 1613 5140
> range(acceleration)
[1]  8.0 24.8
> range(year)
[1] 70 82
```

c.

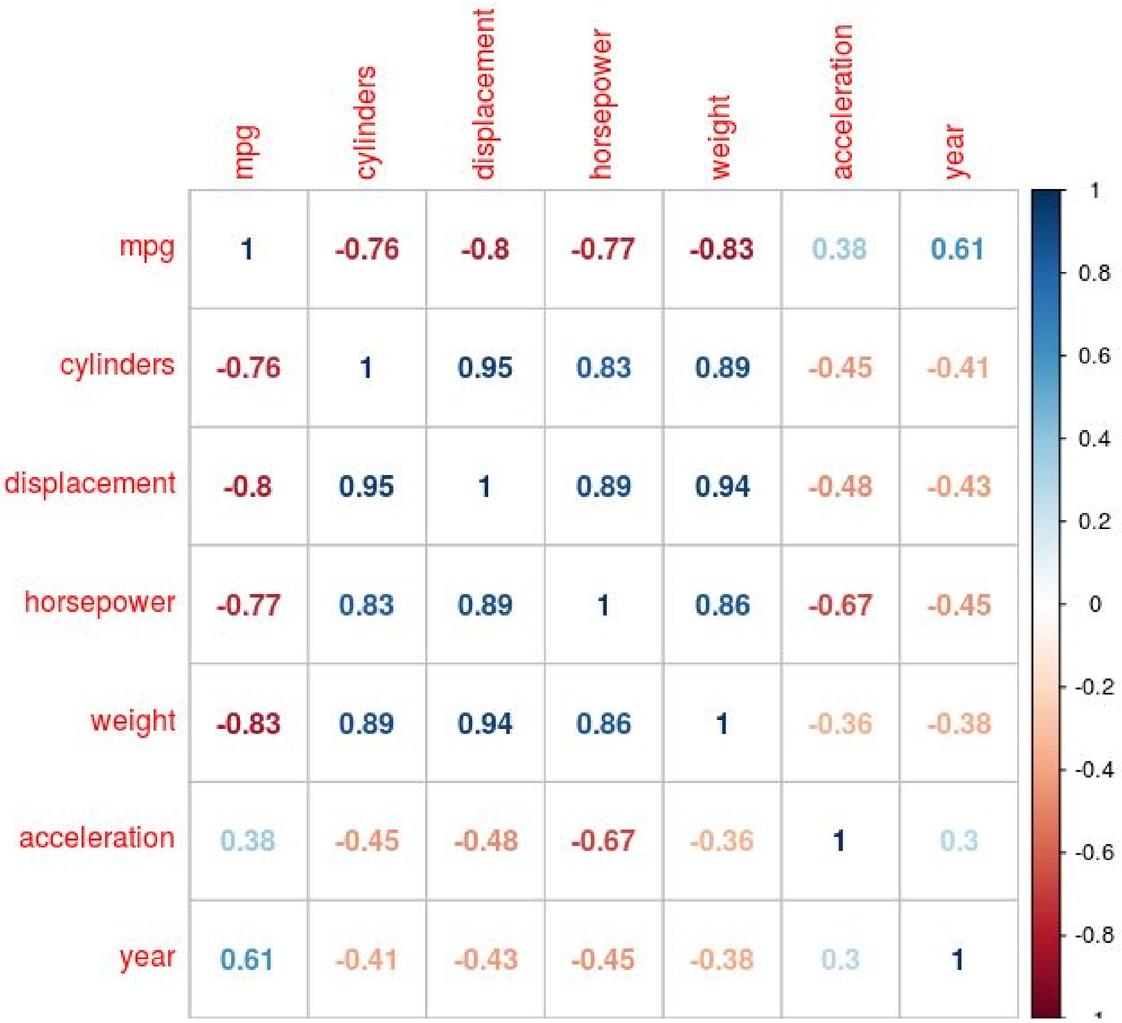
Quantitative	SD	Mean
mpg	7.805007	23.44592
cylinders	1.705783	5.471939
displacement	104.644	194.412
horsepower	38.49116	104.4694
weight	849.4026	2977.584
acceleration	2.758864	15.54133
year	3.683737	75.97959

d.

Quantitative	SD	Mean	Range
mpg	7.867283	24.404430	11.0 - 46.6
cylinders	1.654179	5.373418	3 -8
displacement	99.678367	187.240506	68 - 455
horsepower	35.708853	100.721519	46 - 230
weight	811.300208	2935.971519	1649 - 4997
acceleration	2.693721	15.726899	8.5 - 24.8
year	3.106217	77.145570	70 - 82



e.



	mpg	cylinders	displacement	horsepower	weight	acceleration	year
mpg	1	-0.76	-0.8	-0.77	-0.83	0.38	0.61
cylinders	-0.76	1	0.95	0.83	0.89	-0.45	-0.41
displacement	-0.8	0.95	1	0.89	0.94	-0.48	-0.43
horsepower	-0.77	0.83	0.89	1	0.86	-0.67	-0.45
weight	-0.83	0.89	0.94	0.86	1	-0.36	-0.38
acceleration	0.38	-0.45	-0.48	-0.67	-0.36	1	0.3
year	0.61	-0.41	-0.43	-0.45	-0.38	0.3	1

It appears that we have strong correlations between (displacement and cylinders), (displacement and weight), (displacement and horsepower), (cylinders and weight), (weight and horsepower). We also notice strong negative correlations between (weight and mpg), (displacement and mpg), (horsepower and mpg), and (cylinders and mpg)

f. It appears that the mpg will decrease as cylinders, displacement, horsepower and weight increase.