KAUNAS UNIVERSITY OF TECHNOLOGY

INFORMATICS FACULTY

INTRODUCTION TO ARTIFICIAL INTELLIGENCE

DATA ANALYSIS LAB WORK REPORT

Report author: Jana Strotdresch

Lecturer: doc.dr. Germanas Budnikas

Kaunas, 2020

1. **Select (create) a dataset to perform this and other laboratory works. Your choice must be approved by the tutor.**

Selected Dataset: College

Link: https://vincentarelbundock.github.io/Rdatasets/csv/ISLR/College.csv

Description: Statistics for a large number of US Colleges from the 1995 issue of US News and World Report.

Format: A data frame with 777 observations on the following 18 variables.

`Private`	A factor with levels No and Yes indicating private or public university
`Apps`	Number of applications received
`Accept`	Number of applications accepted
`Enroll`	Number of new students enrolled
`Top10perc`	Pct. new students from top 10% of H.S. class
`Top25perc`	Pct. new students from top 25% of H.S. class
`F.Undergrad`	Number of fulltime undergraduates
`P.Undergrad`	Number of parttime undergraduates
`Outstate`	Out-of-state tuition
`Room.Board`	Room and board costs
`Books`	Estimated book costs
`Personal`	Estimated personal spending
`PhD`	Pct. of faculty with Ph.D.'s
`Terminal`	Pct. of faculty with terminal degree
`S.F.Ratio`	Student/faculty ratio
`perc.alumni`	Pct. alumni who donate
`Expend`	Instructional expenditure per student
`Grad.Rate`	Graduation rate

The columns have the following datatypes:

```
RangeIndex: 777 entries, 0 to 776
Data columns (total 19 columns):
Name           777 non-null object
Private        777 non-null object
Apps           777 non-null int64
Accept         777 non-null int64
Enroll         777 non-null int64
Top10perc      777 non-null int64
Top25perc      777 non-null int64
F.Undergrad    777 non-null int64
P.Undergrad    777 non-null int64
Outstate       777 non-null int64
Room.Board     777 non-null int64
Books          777 non-null int64
Personal       777 non-null int64
PhD            777 non-null int64
Terminal       777 non-null int64
S.F.Ratio      777 non-null float64
perc.alumni    777 non-null int64
Expend         777 non-null int64
Grad.Rate      777 non-null int64
dtypes: float64(1), int64(16), object(2)
memory usage: 115.5+ KB
None
```

2. **For each numeric type attribute calculate:**
   - total number of values,

- percentage of missing values,
- cardinality,
- minimum (min) and maximum (max) values,
- 1st and 3rd quartiles,
- average,
- median,
- Standard deviation.

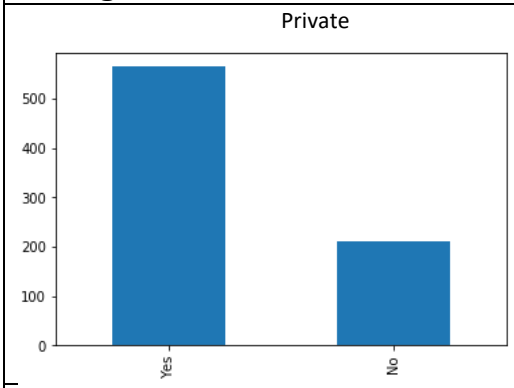
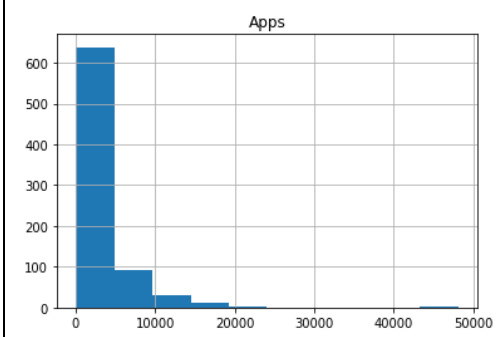
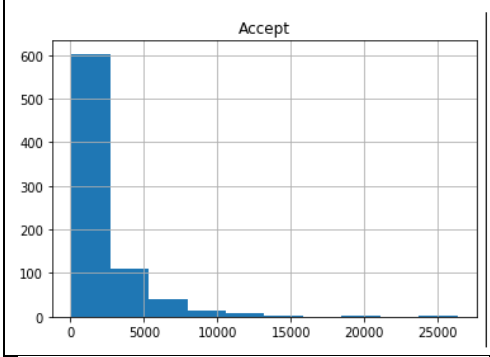
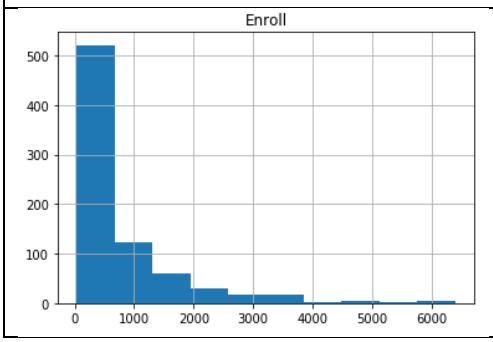
column	TotNrVl	percMiss	cardinality	Min	Max	q1	q3	average	median	StandDeviation
Apps	777	0.0	711	81	48094	776.0	3624.0	3001.6383526383524	1558.0	3870.201484435291
Accept	777	0.0	693	72	26330	604.0	2424.0	2018.8043758043757	1110.0	2451.113970992631
Enroll	777	0.0	581	35	6392	242.0	902.0	779.972972972973	434.0	929.17619013287
Top10perc	777	0.0	82	1	96	15.0	35.0	27.55855855855856	23.0	17.640364385452134
Top25perc	777	0.0	89	9	100	41.0	69.0	55.7966537966538	54.0	19.80477759513136
F.Undergrad	777	0.0	714	139	31643	992.0	4005.0	3699.907335907336	1707.0	4850.42053088738
P.Undergrad	777	0.0	566	1	21836	95.0	967.0	855.2985842985843	353.0	1522.4318872955134
Outstate	777	0.0	640	2340	21700	7320.0	12925.0	10440.66924066924	9990.0	4023.0164841119686
Room.Board	777	0.0	553	1780	8124	3597.0	5050.0	4357.526383526383	4200.0	1096.696415593528
Books	777	0.0	122	96	2340	470.0	600.0	549.3809523809524	500.0	165.10536013709293
Personal	777	0.0	294	250	6800	850.0	1700.0	1340.6422136422136	1200.0	677.0714535905786
PhD	777	0.0	78	8	103	62.0	85.0	72.66023166023166	75.0	16.32815468793933
Terminal	777	0.0	65	24	100	71.0	92.0	79.70270270270271	82.0	14.722358527903365
S.F.Ratio	777	0.0	173	2.5	39.8	11.5	16.5	14.089703989703986	13.6	3.958349135205549
perc.alumni	777	0.0	61	0	64	13.0	31.0	22.743886743886744	21.0	12.39180148937614
Expend	777	0.0	744	3186	56233	6751.0	10830.0	9660.17117117117	8377.0	5221.7684398560905
Grad.Rate	777	0.0	81	10	118	53.0	78.0	65.46332046332046	65.0	17.17770989715541

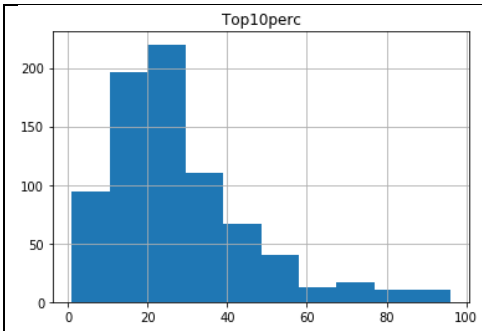
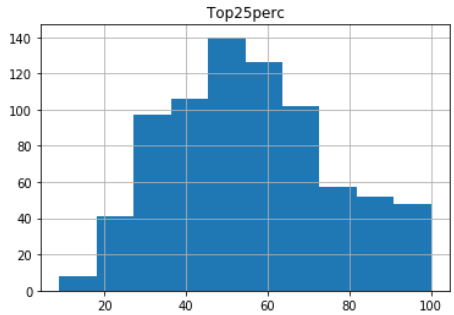
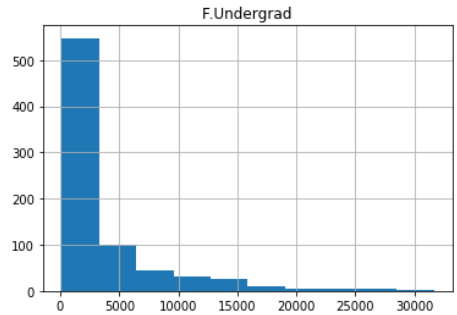
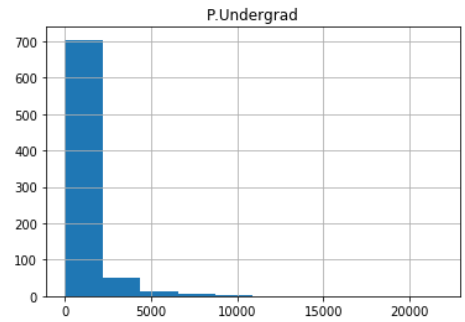
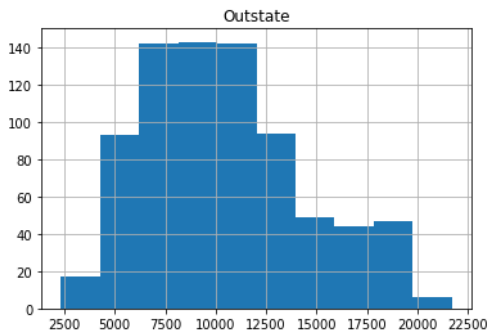
3. For each *category* type attribute calculate:

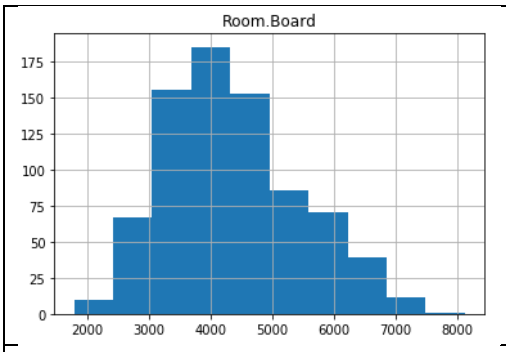
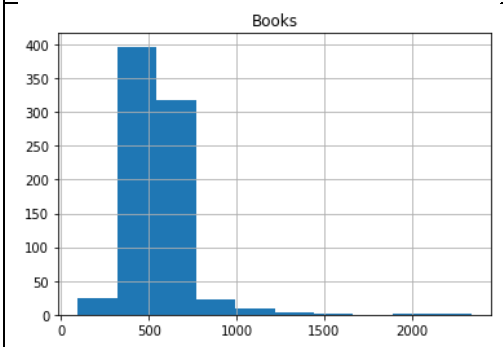
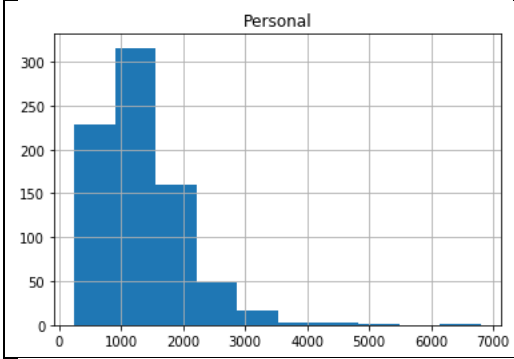
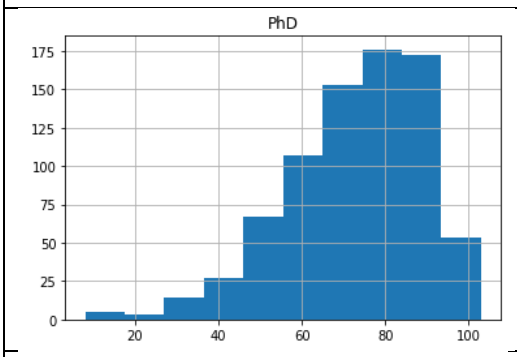
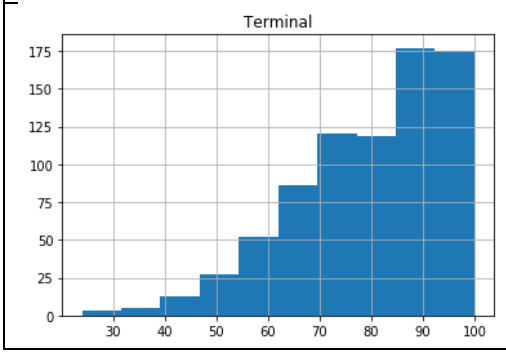
- total number of values,
- percentage of missing values,
- cardinality,
- mode,
- The frequency of the mode
- Percentage value of the mode
- Second mode value (mode 2),
- Frequency value for Mode 2,
- Percentage of Mode 2.

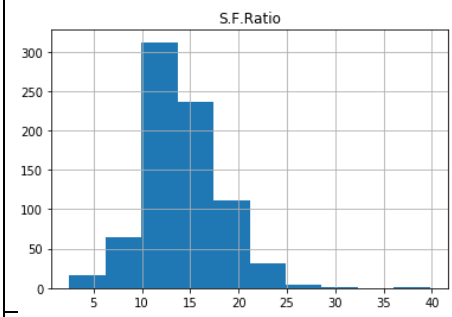
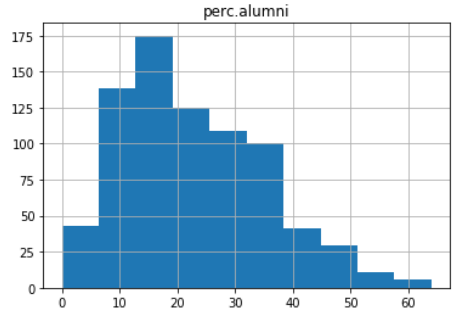
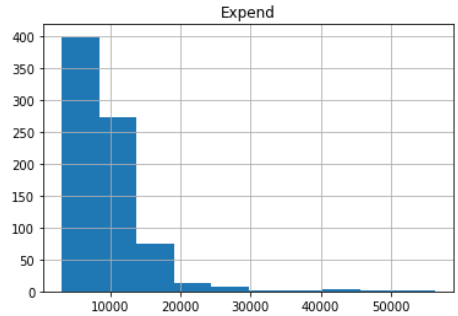
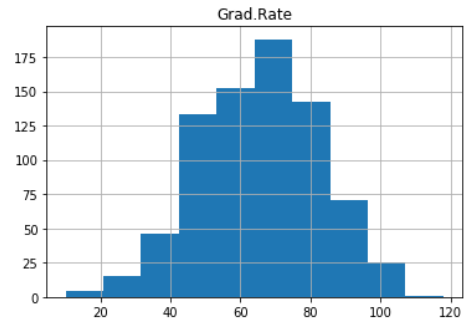
column	TotNrVl	percMiss	cardinality	mode1	freqMode1	percMode1	mode2	freqMode2	percMode2
Private	777	0.0	2	Yes	565	72.71557271557272	No	212	27.28442728442728

4. **Draw histograms of attributes. Provide descriptions of the distribution (eg, normal, exponential, etc.) and what conclusions can be drawn from it.**

Histogram	Description
	Only two categories, there are more private universities than state universities
	Skewed right Outlier between 40000 and 50000 should be removed
	Skewed right Outliers at 20000 and 25000 should be removed
	Skewed right

	Normal distribution, but looks like cut off at 0%.
	Normal, but is cut off at 100%
	Skewed right
	Skewed right
	Normal distribution

	Normal distribution
	Normal distribution
	Normal distribution, but looks like cut off at 0%.
	Skewed left
	Skewed left

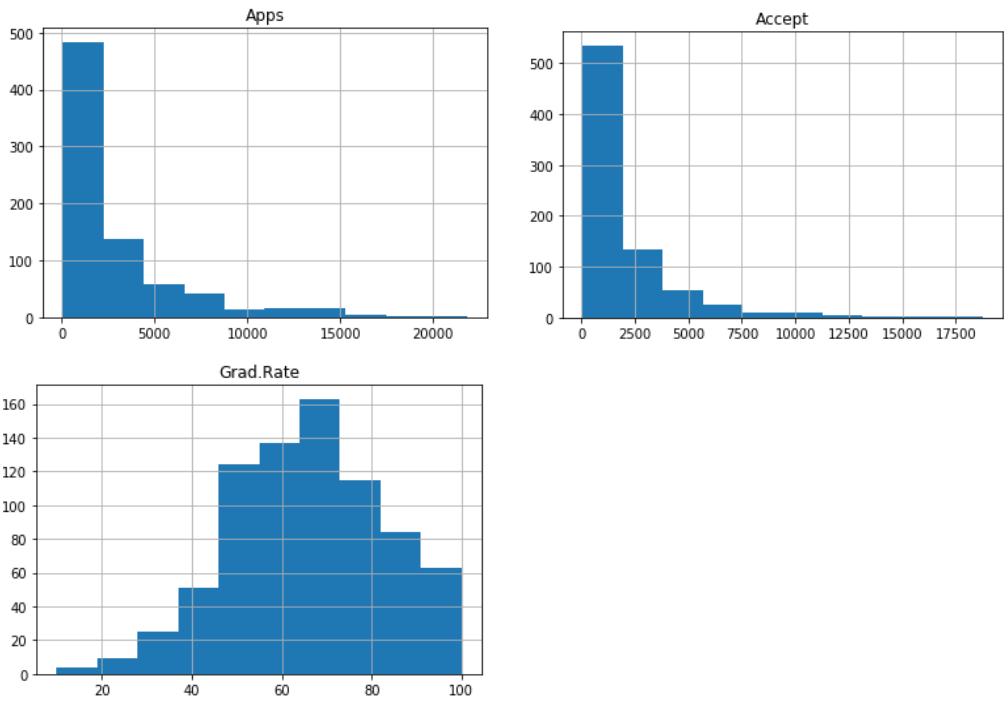
	Normal distribution
	Normal distribution, but also skewed right.
	Skewed right
	Normal The outlier, that has more than 100% should be removed

5. **Identify data quality problems: missing values, cardinality problems, outliers. Provide a plan for resolving these issues, which will be implemented programmatically (e.g., missing values for a categorical attribute based on the attribute estimate of the mode, extreme values being removed or corrected).**

   There are no values missing. So there is to be done.

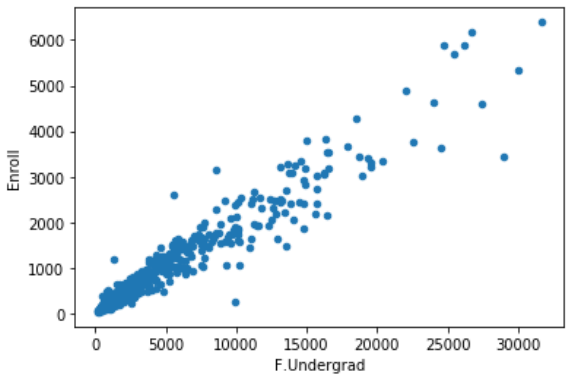
   There are some outliers that should be removed, like I marked in the histogams above.

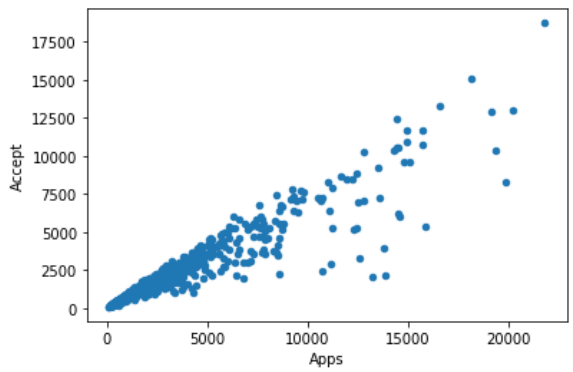
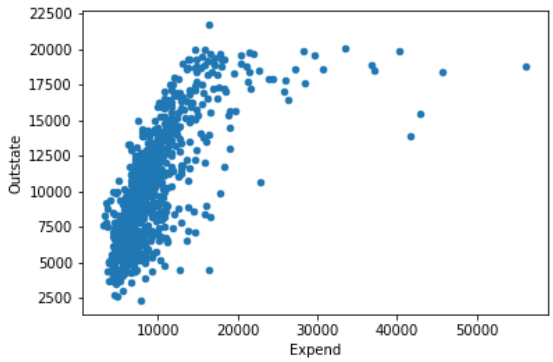
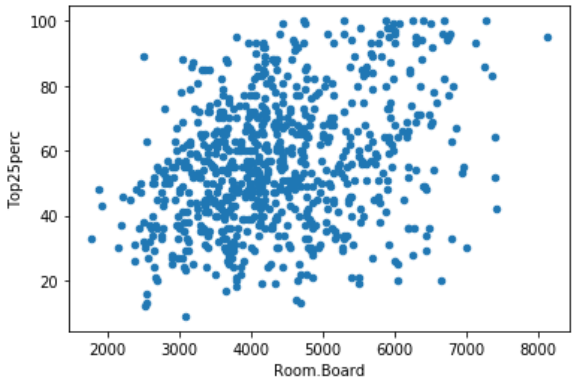
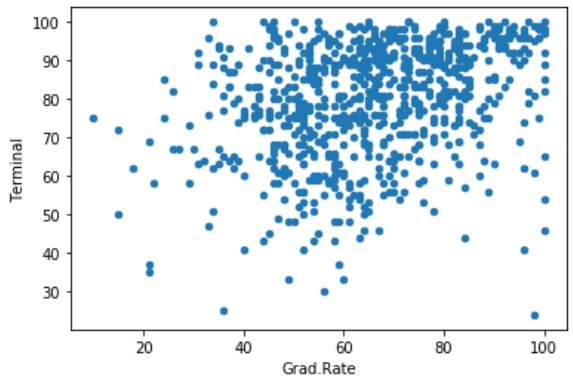
   I have removed all rows where the „Apps“ is higher than 3000. Due to that the outliers of „Accepts“ also distinguished.
   I also have removed all rows that have an „Grad.Rate“ above 100, because a graduation rate of more than 100% doesn‘t make sense.
   The histograms that were problematical before now look like this:

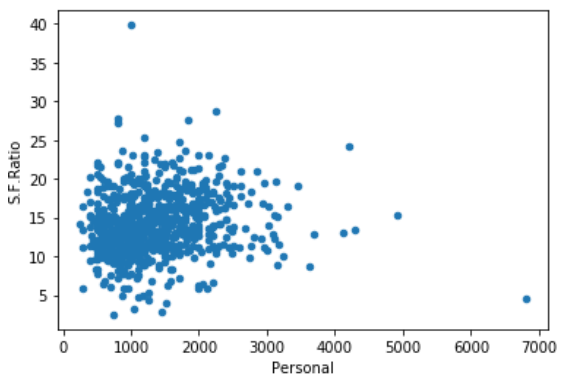


All following tasks are solved with this corrected dataset.

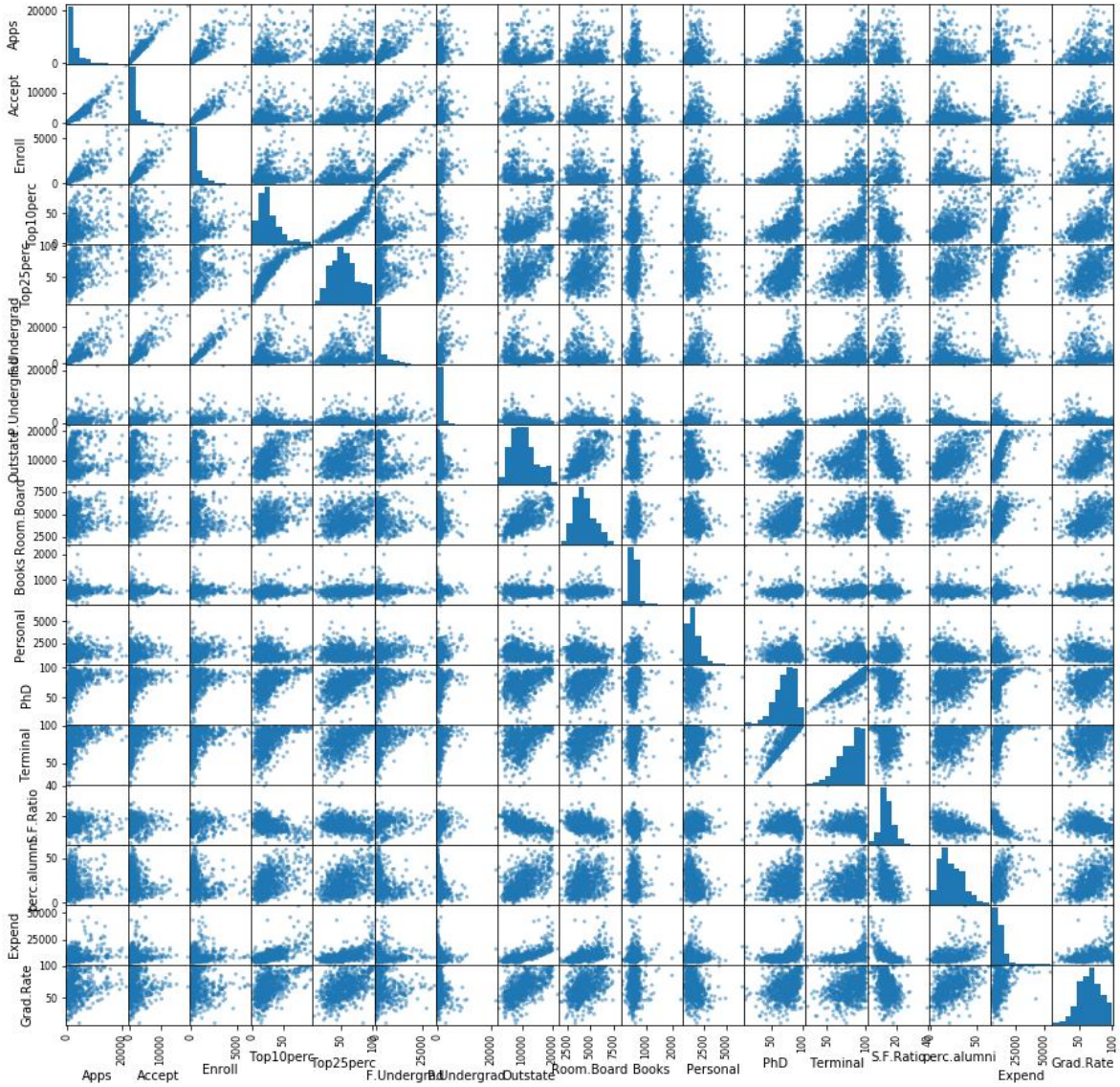
6. **Establish relationships between attributes using visualization techniques**
   - **For numeric type attributes: Using a scatter plot type graph, provide multiple (2-3) examples with strong linear attribute dependency (direct or inverse correlation) and multiple examples with non-correlated (weakly correlated) attributes. Comment on results.**
   - **Provide an SPLOM diagram (Scatter Plot Matrix).**

Scatter Plot	**Description**
	Strong linear correlation: Universities with more fulltime undergraduate students have more newly enrolled students in the measured year.

	Strong linear correlation: The more Applications, the more students the university accepts.
	Linear correlation The higher the universities instructional expenditure per student, the higher are the out-of-state tuition fees. Nevertheless there seems to be a upper frontier of 20000 $.
	No correlation between the Room and board costs and the percentage of Top 25 students.
	No correlation between the graduation rate and the percentage of falculties with terminal degrees.

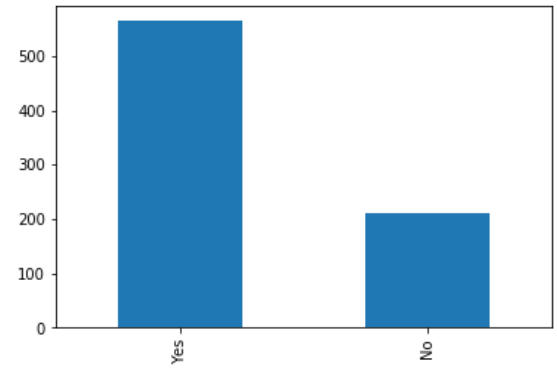
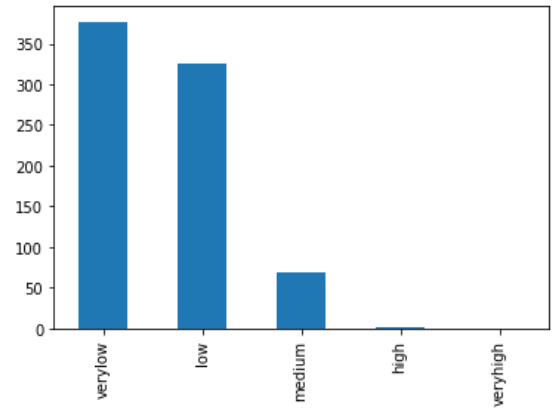
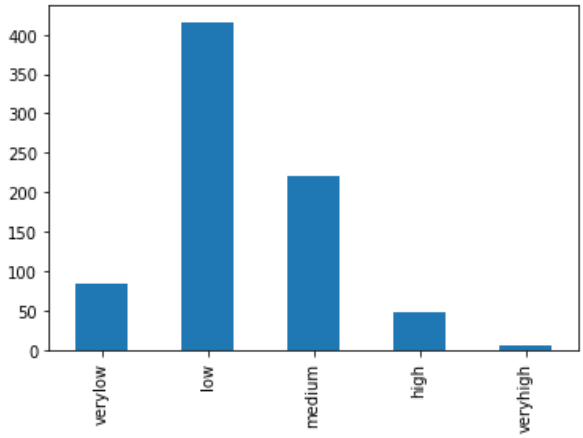
	No correlation between the student/faculty ratio and the estimated personal spendings

SPLOM-Diagram:

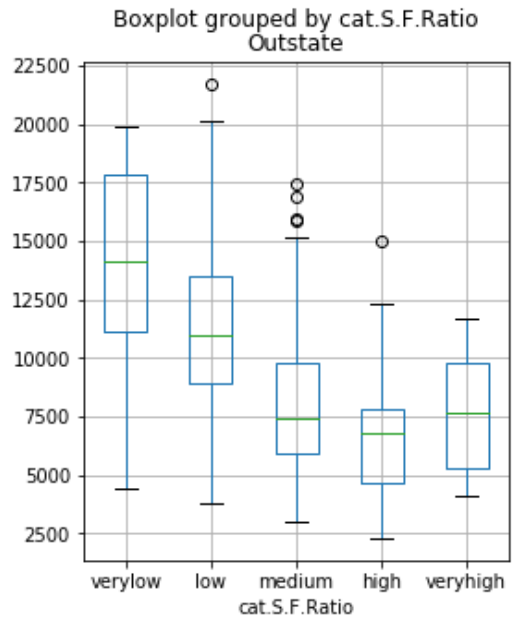
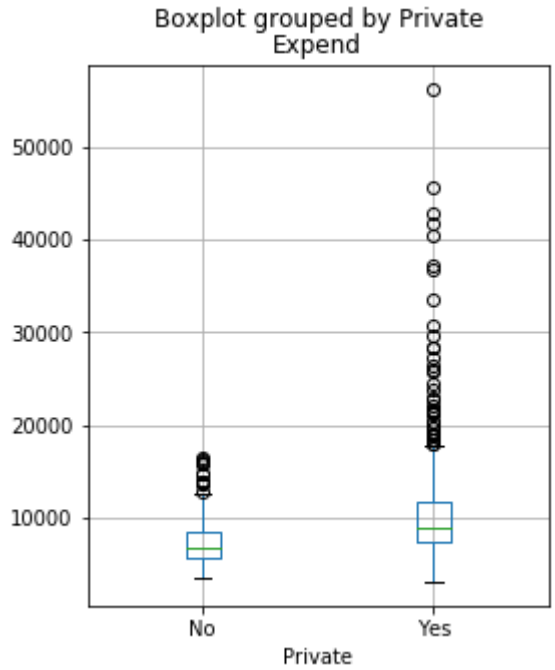


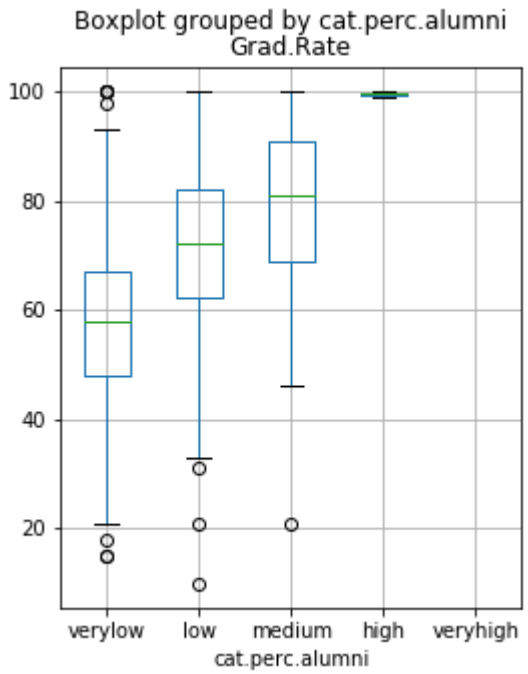
- ***For categorical attributes:*** **Using the bar plot type diagram, give some (2-3) examples of attribute frequency and comment on the results.**

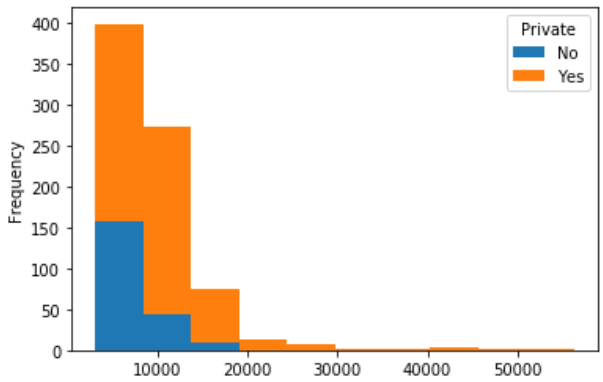
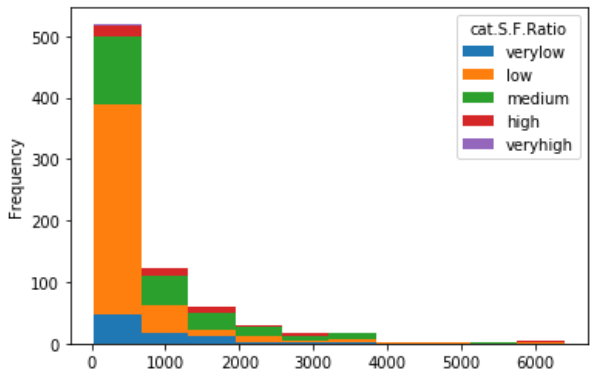
I had only one categorical in my original dataset, which was the column „Private“. For this task I transformed the numericals „S.F.Ratio“ and „perc.alumni“ to categoricals and added them as a new column to my dataset. Both are percentages, but due to the different distributions I made different sized bins for them.

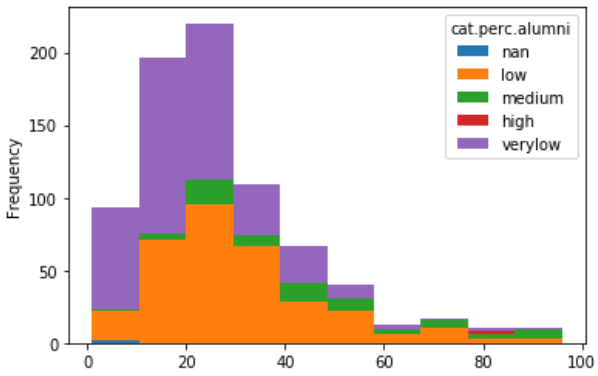
Bar plots	**Legend**	**Comment**
Private 	Yes Private university No State university	Most of Americas universities are private.
Cat.perc.alumni 	Very low 0-20% Low 20-40% Medium 40-60% High 60-80% Very high 80-100%	Only a very low to low rate of alumni donates money tot he university
Cat.S.F.Ratio 	Very low 0-10% Low 10-15% Medium 15-20% High 20-25% Very high 25-50%	Most of the universities have a low student/faculty ratio

- **Provide some (2-3) examples of histograms and box plot diagrams depicting relationships between categorical and numeric type variables.**

Boxplot	**Description**
	The lower the S.F.Ratio, the higher the Out-of-state tuition fees
	Private universities spend in average more money on instructioning every single student than state universities. Also the highest expenditures are from private universities.

	Universities with higher percentage of alumni donating money to the university have higher graduation rates. There are only very few universities with a „high“ percentage of alumni donating, so the boxplot is very small.

Histogram	**Description**
Expend 	State universities have expenditure up to 20000,
Enroll 	Only universities with low numbers of enrollments have a very low students/faculty ratio. The most universities have low enrollment numbers and a low student/faculty ratio

Top10perc 	Those universities with low rates of Top10Percent students have very low percentage of alumni donating.

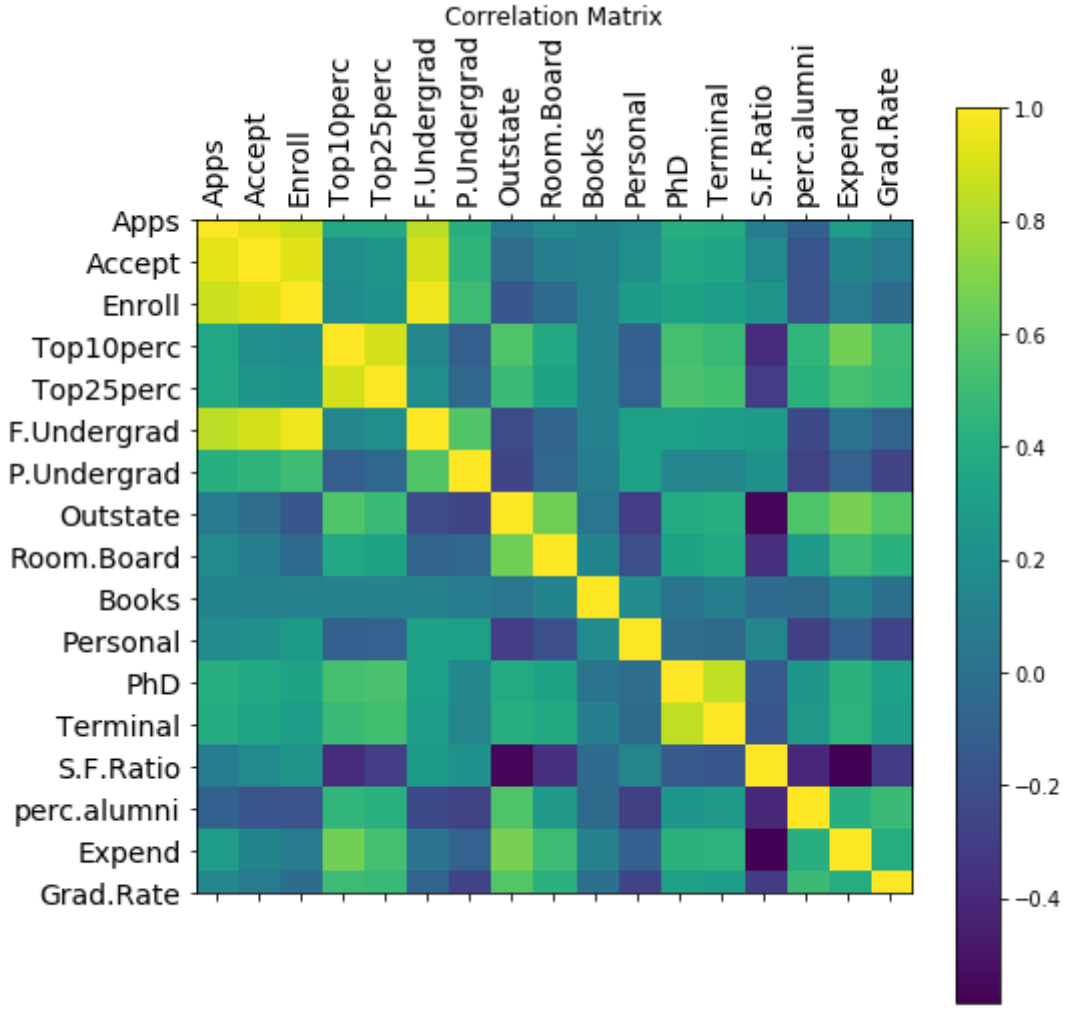
7. **Calculate the covariance and correlation values between continuous attributes and graphically represent the correlation matrix. Comments on the results.**

Covariance:

	Apps	Accept	...	Expend	Grad.Rate
Apps	1.238569e+07	7.553129e+06	...	5.214547e+06	9047.230824
Accept	7.553129e+06	5.256187e+06	...	1.578470e+06	2380.382371
Enroll	2.835244e+06	1.964472e+06	...	3.075825e+05	-396.456973
Top10perc	2.272180e+04	8.111825e+03	...	6.098065e+04	151.514695
Top25perc	2.569304e+04	1.135353e+04	...	5.461029e+04	163.856464
F.Undergrad	1.429975e+07	9.868722e+06	...	4.482188e+05	-6663.127557
P.Undergrad	2.187022e+06	1.562651e+06	...	-6.712066e+05	-6723.984096
Outstate	9.610188e+05	-1.573198e+05	...	1.417025e+07	39699.029391
Room.Board	6.785395e+05	2.317949e+05	...	2.881548e+06	7987.383721
Books	7.666039e+04	4.154402e+04	...	9.714390e+04	-2.535642
Personal	4.315506e+05	3.149689e+05	...	-3.498274e+05	-3093.593065
PhD	2.380092e+04	1.382435e+04	...	3.684661e+04	88.961949
Terminal	2.025332e+04	1.179359e+04	...	3.372156e+04	75.403201
S.F.Ratio	1.152988e+03	1.543695e+03	...	-1.210389e+04	-21.004294
perc.alumni	-4.116695e+03	-4.749157e+03	...	2.709573e+04	105.006568
Expend	5.214547e+06	1.578470e+06	...	2.733149e+07	35224.687355
Grad.Rate	9.047231e+03	2.380382e+03	...	3.522469e+04	292.091356

Correlation:

	Apps	Accept	Enroll	...	perc.alumni	Expend	Grad.Rate
Apps	1.000000	0.936120	0.875145	...	-0.094283	0.283416	0.150417
Accept	0.936120	1.000000	0.930808	...	-0.166965	0.131695	0.060751
Enroll	0.875145	0.930808	1.000000	...	-0.181230	0.063912	-0.025199
Top10perc	0.365838	0.200489	0.180505	...	0.455805	0.660947	0.502345
Top25perc	0.368739	0.250126	0.223005	...	0.418727	0.527602	0.484248
F.Undergrad	0.844071	0.894203	0.964091	...	-0.230250	0.017810	-0.080990
P.Undergrad	0.408667	0.448232	0.509759	...	-0.280938	-0.084431	-0.258728
Outstate	0.067817	-0.017042	-0.153336	...	0.566177	0.673151	0.576882
Room.Board	0.175614	0.092090	-0.042385	...	0.272709	0.502039	0.425685
Books	0.131832	0.109668	0.109608	...	-0.039812	0.112458	-0.000898
Personal	0.181168	0.202975	0.278765	...	-0.286429	-0.098863	-0.267432
PhD	0.416551	0.371401	0.330651	...	0.250299	0.434111	0.320612
Terminal	0.391912	0.350318	0.306526	...	0.267965	0.439267	0.300457
S.F.Ratio	0.082759	0.170089	0.232928	...	-0.402904	-0.584849	-0.310455
perc.alumni	-0.094283	-0.166965	-0.181230	...	1.000000	0.417747	0.495224
Expend	0.283416	0.131695	0.063912	...	0.417747	1.000000	0.394236
Grad.Rate	0.150417	0.060751	-0.025199	...	0.495224	0.394236	1.000000



8. Perform data normalization.

I converted all values in each column to values between 0 and 1, except for the column of „Name“. In this task the categoricals have already been converted to numericals, like it sais in task 9.

	Private	Apps	...	Grad.Rate	cat.perc.alumni
count	775.000000	775.000000	...	775.000000	773.000000
mean	0.727742	0.131720	...	0.615341	0.201811
std	0.445409	0.162009	...	0.189896	0.218811
min	0.000000	0.000000	...	0.000000	0.000000
25%	0.000000	0.031994	...	0.477778	0.000000
50%	1.000000	0.067946	...	0.611111	0.333333
75%	1.000000	0.161580	...	0.755556	0.333333
max	1.000000	1.000000	...	1.000000	1.000000

9. Convert categorical variables to numeric type variables.

For this task I transformed the the „S.F.Ratio“ again to a categorical like in task 6, so that I have at least two categoricals.

I wrote a fuction to transform the categoricals „Private“ and „cat.perc.alumni“ to numeric values and applied it to the dataset.

```
                              Name  Private  ...  Grad.Rate  cat.perc.alumni
0      Abilene Christian University        1  ...         60                0
1                Adelphi University        1  ...         56                0
2                    Adrian College        1  ...         54                1
3               Agnes Scott College        1  ...         59                1
4         Alaska Pacific University        1  ...         15                0
..                              ...      ...  ...        ...              ...
772         Worcester State College        0  ...         40                0
773               Xavier University        1  ...         83                1
774  Xavier University of Louisiana        1  ...         49                0
775                 Yale University        1  ...         99                2
776    York College of Pennsylvania        1  ...         99                1
```