# *Assignment 1*

## *Task 25)*

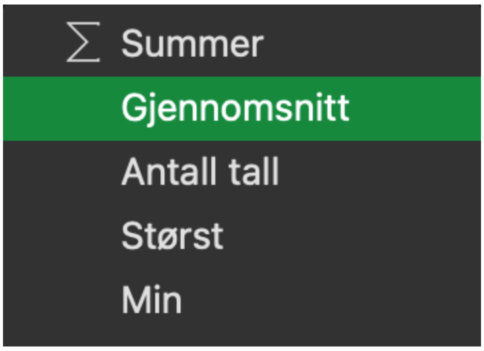
a) There are 7 variables in the data set; Make, Model, Overall Score, Recommended, Owner Satisfaction, Overall Miles per Gallon and Acceleration.

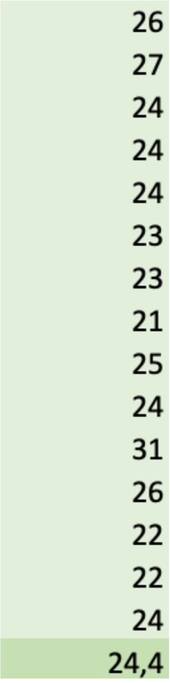
b) Categorical variables: Make, Model, Recommended, Owner Satisfaction, Overall Score.

Quantitative variables: Overall Miles per Gallon, Acceleration.

c) $\frac{7}{15} * 100\% = 46{,}67\%$ of the 15 vehicles are recommended.

d) The average of the overall miles per gallon across all 15 vehicles is 24.4.



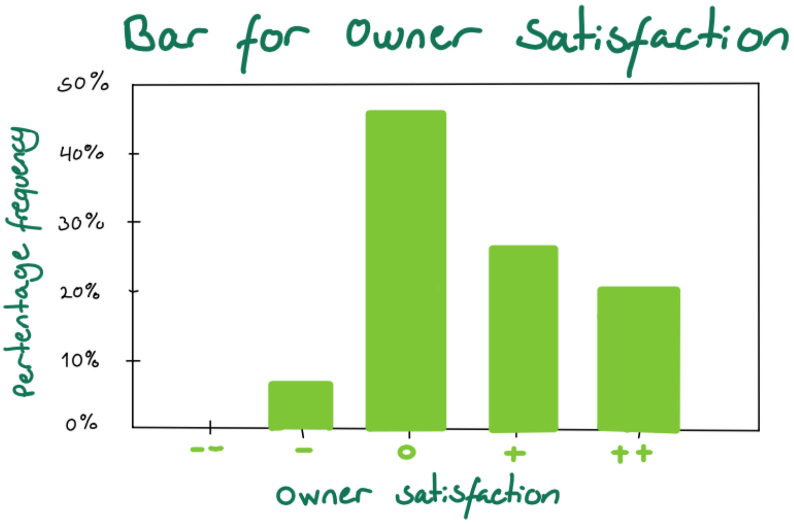


| |
|---:|
| 26 |
| 27 |
| 24 |
| 24 |
| 24 |
| 23 |
| 23 |
| 21 |
| 25 |
| 24 |
| 31 |
| 26 |
| 22 |
| 22 |
| 24 |
| 24,4 |

e) See python file

## Perctentage frequency

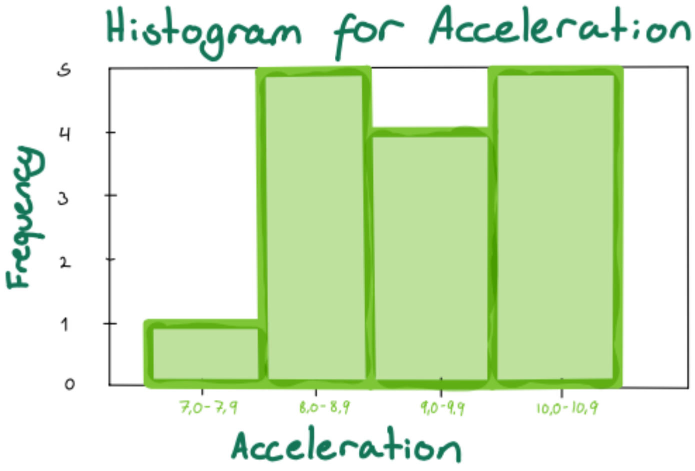
"--" = 0 -> 0/15·100 = 0%
"-" = 1 -> 1/15·100 = 6,66%
"0" = 7 -> 7/15·100 = 46,66%
"+" = 4 -> 4/15·100 = 26,68%
"++" = 3 -> 3/15·100 = 20%



f) See python file

## Frequency distribution

7,0 - 7,9 = 1 , 8,0 - 8,9 = 5
9,0 -> 9,9 = 4 , 10,0 - 10,9 = 5



*Task 48)*

a) See python file

## Frequency

I used a for-loop in python to calculate the frequency of complaints by industry

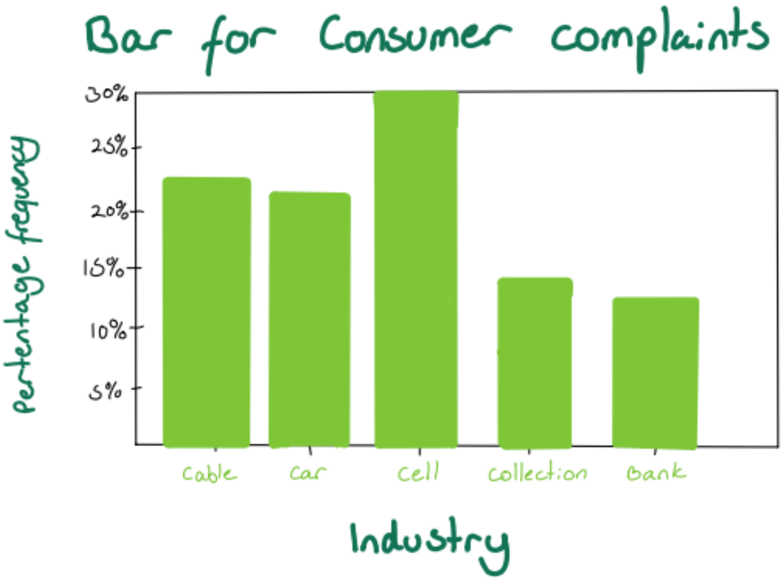
Result:

| | |
|---|---|
| Cable | 44 |
| Car | 42 |
| Cell | 60 |
| Collection | 28 |
| Bank | 26 |

## Percent frequency

| | |
|---|---|
| Cable: | 44/200 · 100 = 22% |
| Car: | 42/200 · 100 = 21% |
| Cell: | 60/200 · 100 = 30% |
| Collection: | 28/200 · 100 = 14% |
| Bank: | 26/200 · 100 = 13% |

b) See python file

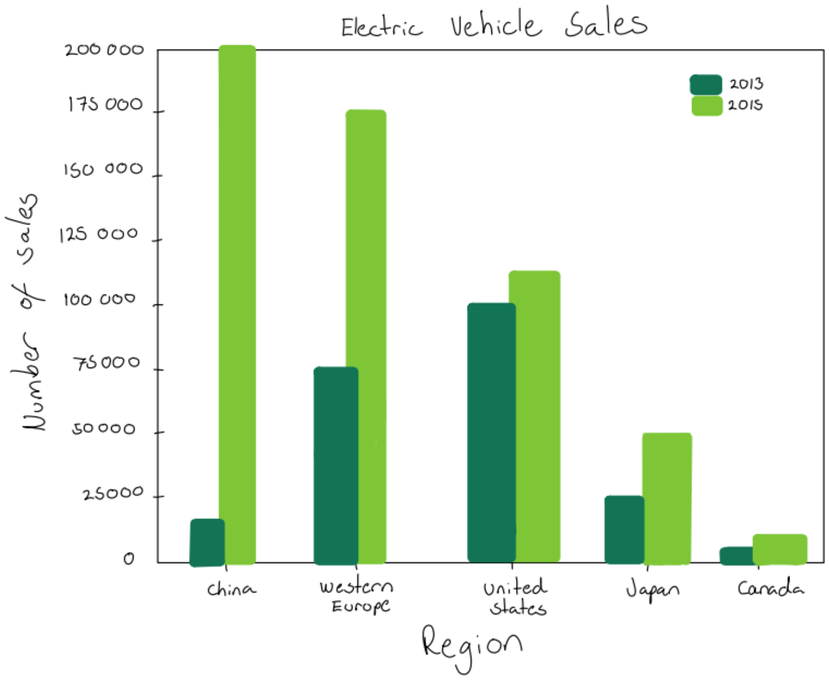


c) Cellular phone providers had the highest number of complaints.

d) Banks and Collection agencies have the lowest number of complaints. Overall the complaints are spread evenly, however, cellular phone providers have the highest percentage.

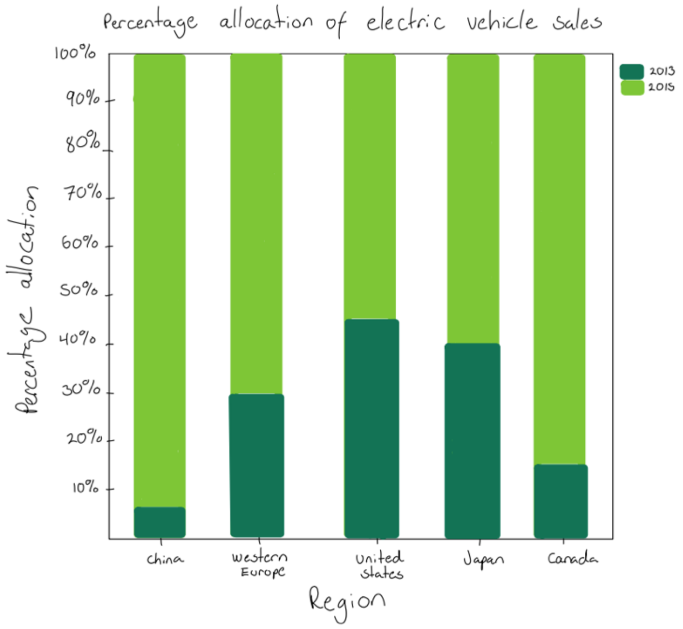
## *Task 57)*

a) See python file



We can see that electric car sales have increased in all regions, regardless of how large the sales were in advance. The increase is particularly large in the number of figures in China and Western Europe. However, Japan and Canada also have a large increase, if we look at relative terms. Still, China has the largest increase by far, both in number of figures and in relative terms.

b) See python file



c) I think the display in (a) is more insightful because it also displays the number of electric plug-in vehicle sales. The number of sales in China is far greater than the number in Canada, for example. The display in (b) does not display the large differences in sales between the regions. The display in (b) only display the relative distribution, while the display in (a) provides an insight in both relative and actual distribution.

## *Task 62)*

a)

$$\begin{array}{cccccccccc} 0 & 0 & 1 & 1 & 1 & 1 & 1 & 2 & 3 & 3 \\ 3 & 3 & 3 & 4 & 4 & 5 & 5 & 6 & 6 & 7 \end{array}$$

**Mean:** $\bar{X} = \frac{0+0+1+1+1+1+1+2+3+3+3+3+3+4+4+5+5+6+6+7}{20} = 2,95$

**Median:** $\frac{X_{10} + X_{11}}{2} = \frac{3+3}{2} = 3$

b)

**First quartile:** $L_{25} = \frac{20+1}{4} = 5,25, \quad Q_1 = 1 + 0,25(1-1) = 1$

**Third quartile:** $L_{75} = \frac{3}{4} \cdot 20+1 = 15,75, \quad Q_3 = 4 + 0,75(5-4) = 4,75$

c)

**Range:** Largest value - Smallest value = $7 - 0 = 7$

**Inter-quartile range:** $Q_3 - Q_1 = 4{,}75 - 1 = 3{,}75$

d)

**Sample variance:**

$$S^2 = \frac{2(0-2{,}95)^2 + 5(1-2.95)^2 + (2-2{,}95)^2 + 5(3-2{,}95)^2 + 2(4-2{,}95)^2 + 2(5-2{,}95)^2 + 2(6-2{,}95)^2 + (7-2{,}95)^2}{20-1}$$

$S^2 = 4{,}37$

**Standard deviation:** $S = \sqrt{4{,}37} = 2{,}09$

e) The skewness measure for there data is 0.34. This indicates that the distribution is *relatively* close to symmetric "normal" distribution. The distribution is also moderately skewed right. This is due to a large number of "1" and a small number of "6" and "7", so that one side "weighs" more. However, since large parts of the dataset is close to the median, the distribution is close to a normal distribution.

f) Lower limit = 1 - 0.5(3.75) = -0.88

Upper limit = 4.75 + 0.5(3.75) = 6.63

We see that the data contains one outlier, 7.

## *Task 70)*

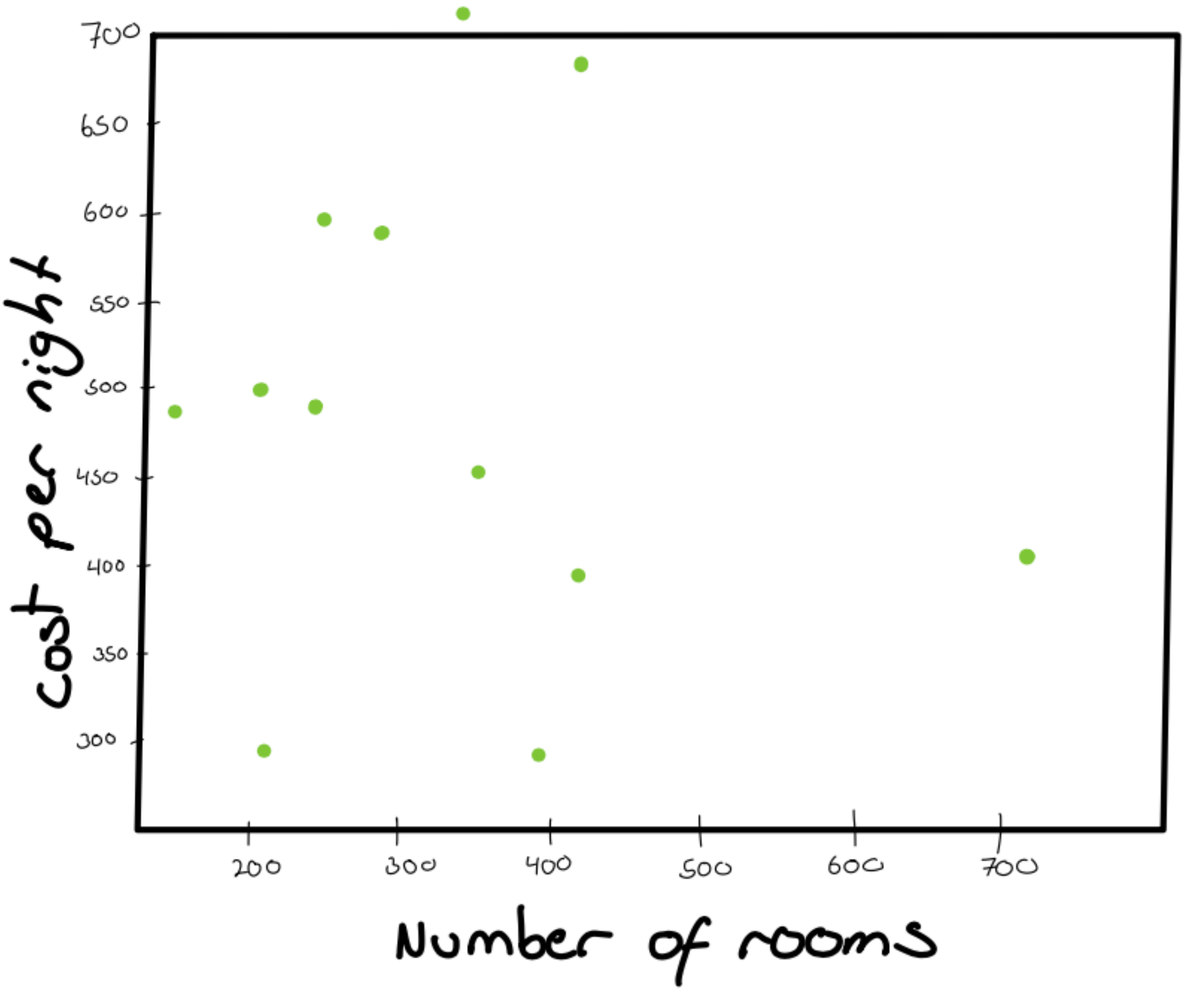
a)

**Mean:** $$\frac{220+727+285+273+145+213+398+343+250+414+400+700}{12} = 364$$

b)

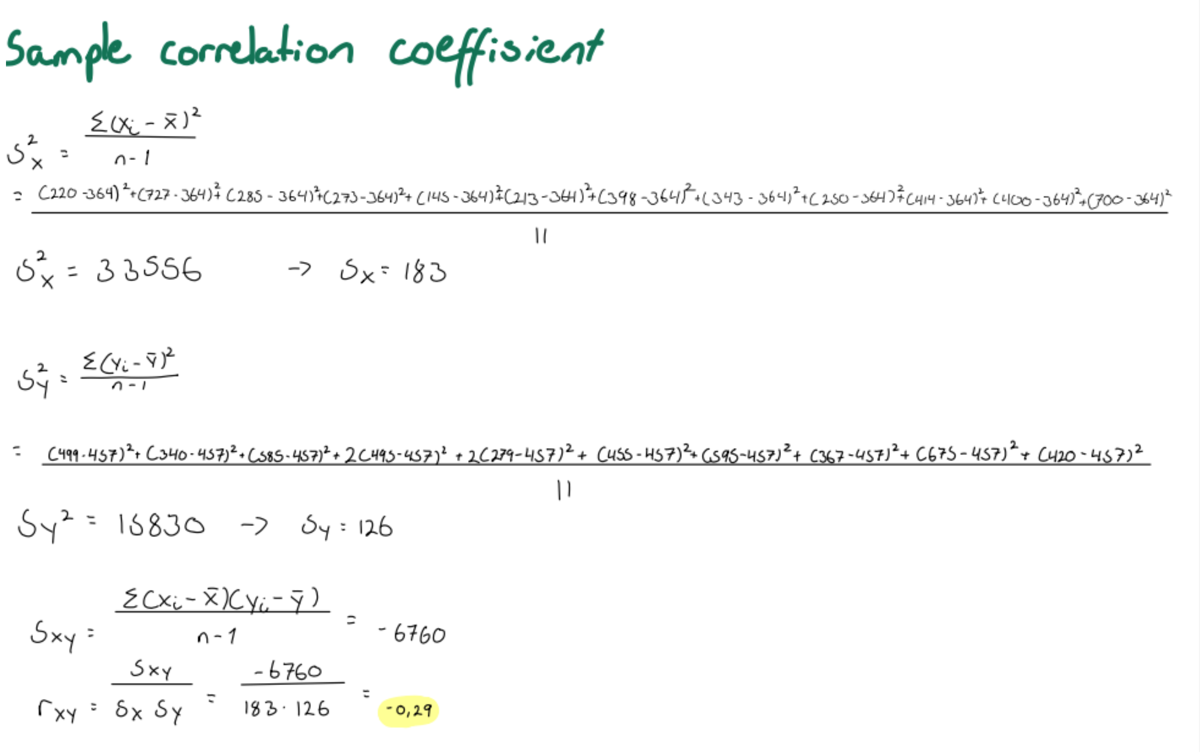
$$\text{Mean: } \frac{499+340+585+495+495+279+279+455+595+367+675+420}{12} = 457$$

c) See python file



There does not seem to be a relationship between the number of rooms and the cost per night. As we can see, the dots are scattered across the chart and are far from the center. This does not give an impression of whether there is a positive or negative relationship between the two variables.

d)



## Sample correlation coeffisient

$$S_x^2 = \frac{\Sigma(x_i - \bar{x})^2}{n-1}$$

$$= \frac{(220-364)^2+(727-364)^2+(285-364)^2+(273-364)^2+(145-364)^2+(213-364)^2+(398-364)^2+(343-364)^2+(250-364)^2+(414-364)^2+(400-364)^2+(700-364)^2}{11}$$

$S_x^2 = 33556 \quad \rightarrow \quad S_x = 183$

$$S_y^2 = \frac{\Sigma(y_i - \bar{y})^2}{n-1}$$

$$= \frac{(499-457)^2+(340-457)^2+(585-457)^2+2(495-457)^2+2(279-457)^2+(455-457)^2+(595-457)^2+(367-457)^2+(675-457)^2+(420-457)^2}{11}$$

$S_y^2 = 15830 \quad \rightarrow \quad S_y = 126$

$$S_{xy} = \frac{\Sigma(x_i - \bar{x})(y_i - \bar{y})}{n-1} = -6760$$

$$r_{xy} = \frac{S_{xy}}{S_x S_y} = \frac{-6760}{183 \cdot 126} = -0{,}29$$

We can see that the sample correlation coefficient is close to zero, which means that there is no linear relationship between the number of rooms and the cost per night for a double room. Since the number is negative, there is a slightly negative relationship between the two variables, but this relationship is very weak. I think that it is reasonable that the number of rooms is not related to the price, because there are many factors that can affect the price of a hotel room. Factors such as location, brand, facilities and quality probably play a much greater role for the price than the size of the hotel.

The slightly negative relationship is probably due to the fact that the operating cost per hotel room is lower the larger the hotel is, because many of the fixed costs are spread over several rooms. Therefore the hotel may charge a lower price.

*Task 74)*

a)

## Mean speed:

I use the median as an estimate for the speed intervals

$$\bar{X} = \frac{10 \cdot (4\overset{47}{5-}49) + 40 \cdot (5\overset{52}{0-}54) + 150(5\overset{57}{5-}59) + 175(6\overset{62}{0-}64) + 75(6\overset{67}{5-}69) + 15(7\overset{72}{0-}74) + 10(7\overset{77}{5-}79)}{475}$$

$\bar{X} = 60,68$ -> the $(60-64)$ interval

b)

## Variance and standard deviation:

$$s^2 = \frac{10(47-61)^2 + 40(52-61)^2 + 150(57-61)^2 + 175(62-61)^2 + 75(67-61)^2 + 15(72-61)^2 + 10(77)^2}{474}$$

$s^2 = 31,33$ -> $s = 5,6$