London
Business
School

Online Exam

Full Course Code	MAM2022 AM01 AUT21
Course Title	Applied Statistics
Faculty	Kostis Christodoulou
Date of Exam	24 September 2021

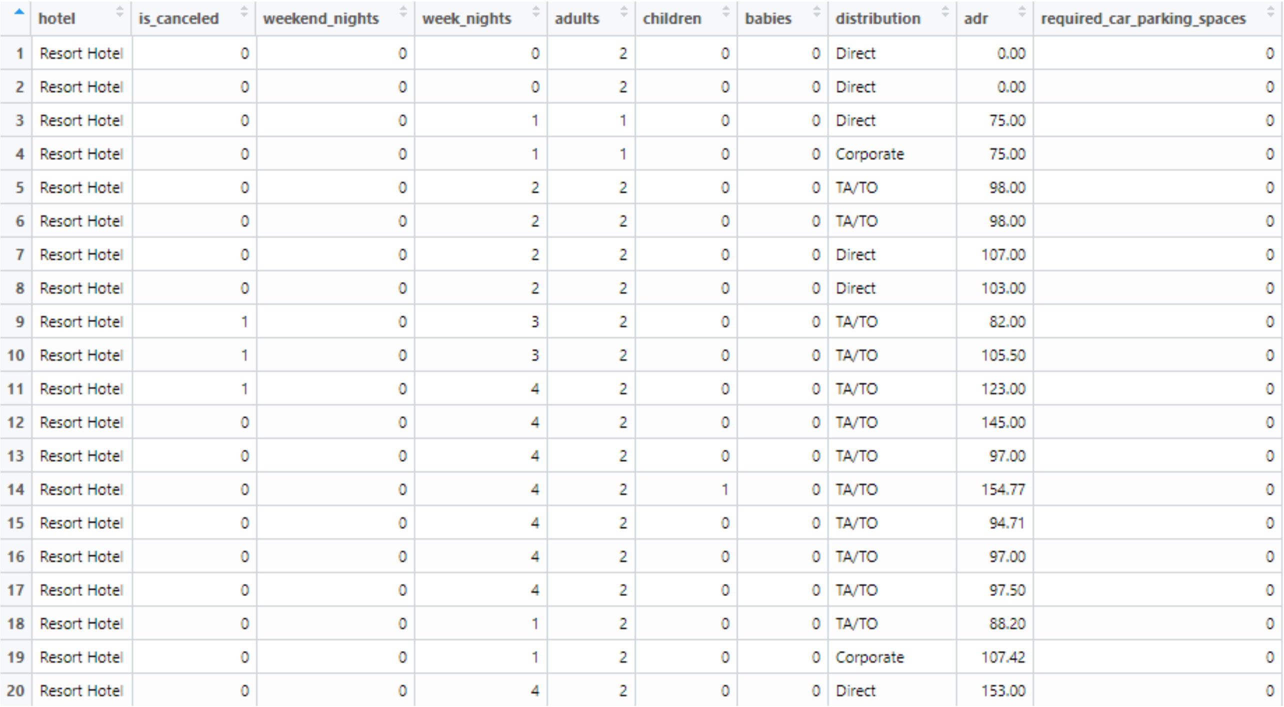
STUDENT TO COMPLETE BEFORE STARTING EXAM							
LBS Number	3	5	0	1	8	1	2
Stream (if applicable)	-						

By downloading this exam paper, you have:

- Read and understood the **Online Exams Policy** and **Exam Instructions**.
- Agreed to complete the exam **individually** without discussion or collaboration with others.
- Agreed to upload and submit the file(s) on Canvas before the final 15 minutes of the exam, and only use the **final 15 minutes** to resolve any **technical issues.**
- Understood that submissions received **after** the deadline will be subject to a late submission penalty as per section 5 of the Online Exams Policy.

Question 1 (25 points)

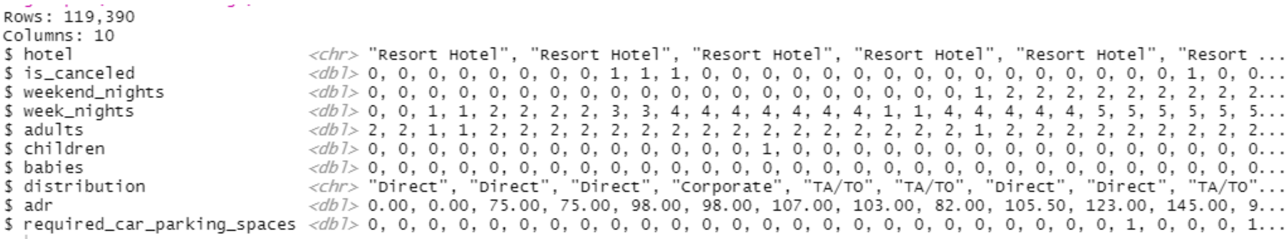
You are presented with data on ***hotel_bookings***, with data on two hotels, a Resort and a City Hotel. Here is what the dataframe looks like



	hotel	is_canceled	weekend_nights	week_nights	adults	children	babies	distribution	adr	required_car_parking_spaces
1	Resort Hotel	0	0	0	2	0	0	Direct	0.00	0
2	Resort Hotel	0	0	0	2	0	0	Direct	0.00	0
3	Resort Hotel	0	0	1	1	0	0	Direct	75.00	0
4	Resort Hotel	0	0	1	1	0	0	Corporate	75.00	0
5	Resort Hotel	0	0	2	2	0	0	TA/TO	98.00	0
6	Resort Hotel	0	0	2	2	0	0	TA/TO	98.00	0
7	Resort Hotel	0	0	2	2	0	0	Direct	107.00	0
8	Resort Hotel	0	0	2	2	0	0	Direct	103.00	0
9	Resort Hotel	1	0	3	2	0	0	TA/TO	82.00	0
10	Resort Hotel	1	0	3	2	0	0	TA/TO	105.50	0
11	Resort Hotel	1	0	4	2	0	0	TA/TO	123.00	0
12	Resort Hotel	0	0	4	2	0	0	TA/TO	145.00	0
13	Resort Hotel	0	0	4	2	0	0	TA/TO	97.00	0
14	Resort Hotel	0	0	4	2	1	0	TA/TO	154.77	0
15	Resort Hotel	0	0	4	2	0	0	TA/TO	94.71	0
16	Resort Hotel	0	0	4	2	0	0	TA/TO	97.00	0
17	Resort Hotel	0	0	4	2	0	0	TA/TO	97.50	0
18	Resort Hotel	0	0	1	2	0	0	TA/TO	88.20	0
19	Resort Hotel	0	0	1	2	0	0	Corporate	107.42	0
20	Resort Hotel	0	0	4	2	0	0	Direct	153.00	0

The variables types are shown below and the two that are not obvious are

is_canceled: 0 if the booking was not cancelled, 1 if it was cancelled
adr: Average Daily Rate, in Euros



```
Rows: 119,390
Columns: 10
$ hotel                       <chr> "Resort Hotel", "Resort Hotel", "Resort Hotel", "Resort Hotel", "Resort Hotel", "Resort ...
$ is_canceled                 <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0...
$ weekend_nights              <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2...
$ week_nights                 <dbl> 0, 0, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 1, 1, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5...
$ adults                      <dbl> 2, 2, 1, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 1, 2, 2, 2, 2, 2, 2, 2, 2, 2...
$ children                    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ babies                      <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
$ distribution                <chr> "Direct", "Direct", "Direct", "Corporate", "TA/TO", "TA/TO", "Direct", "Direct", "TA/TO"...
$ adr                         <dbl> 0.00, 0.00, 75.00, 75.00, 98.00, 98.00, 107.00, 103.00, 82.00, 105.50, 123.00, 145.00, 9...
$ required_car_parking_spaces <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1...
```

Use dplyr commands that will produce the following. You don't need to calculate the output table, just write the code that would produce it.

Part a (5 pts) Which distribution channel (***distribution***) generates the greatest number of bookings in terms of volume? Calculate both counts and percentages and sort your table in descending order

```
hotel_bookings %>%
        group_by(distribution)%>%
        summarise(count = n()) %>%
        mutate( percentage = count/n()) %>%
        arrange(-count)
```

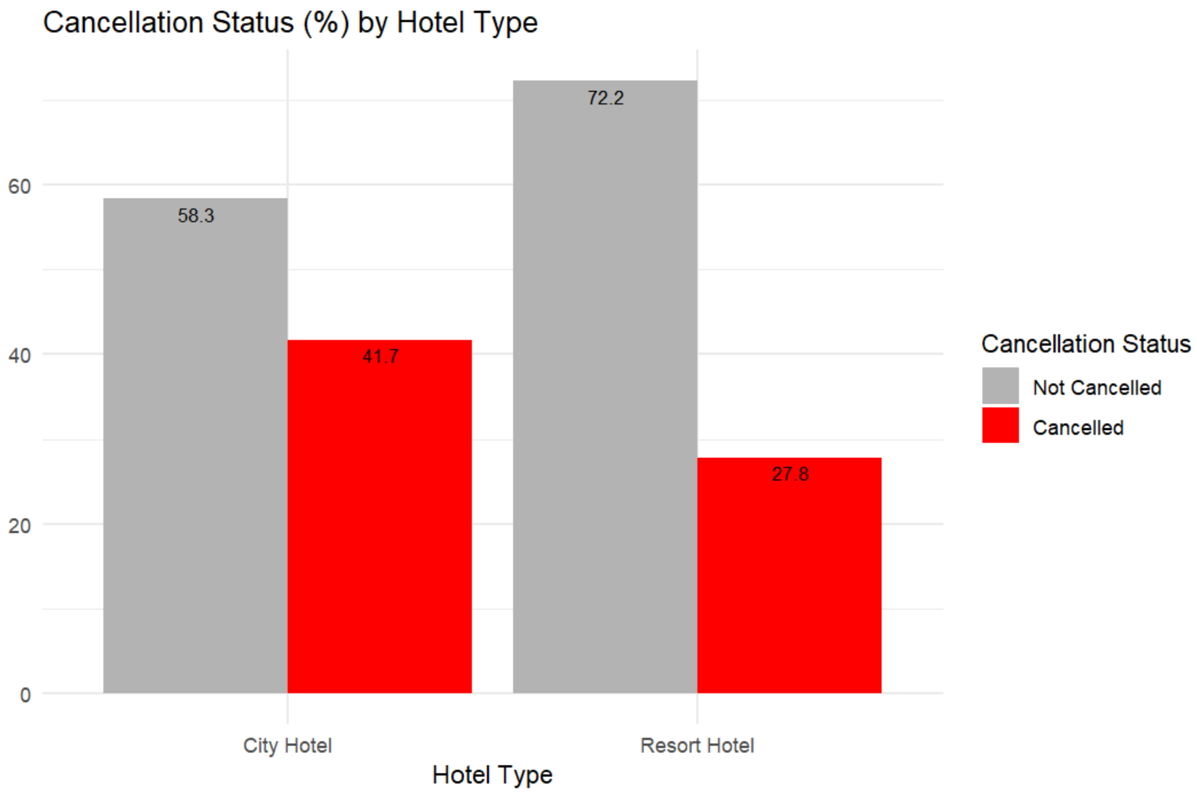
Part b (5 pts) Which distribution channel (***distribution***) generates the greatest revenue? Calculate both the total revenue each distribution channel generated and its percentage contribution to the total revenue. Use the average daily revenue (***adr***) and apply that to both week_nights and weekend_nights.

```
hotel_bookings %>%
        mutate (total_nights = week_nights + weekend_nights,
                           revenue = total_nights * adr) %>%
        group_by (distribution) %>%
        summarise (total_revenue = sum(revenue))  %>%
        mutate (revenue_percentage = total_revenue / sum(total_revenue)) %>%
        slice_max(order_by = revenue_percentage, n=1)
```

Part c (5 pts) For those customers who did stay, what proportion had a booking that involved kids (either *children* or *babies*).

```
proportion_kids <- hotel_bookings %>%
    filter( (babies>0 | children > 0) & adr > 0  ) %>%
    mutate ( kids = babies+children) %>%
    mutate ( have_kids  = ifelse(  ( kids > 0 ) , "Yes" , "No" )) %>%
    group_by(have_kids) %>%
    summarise(count_have = n()) %>%
    mutate(proportion = count_have/sum(count_have))
```

Part d (10 pts) Using tidyverse packages and functions, how you would you create the following plot that looks at cancellation status (***is_cancelled***) by hotel type?



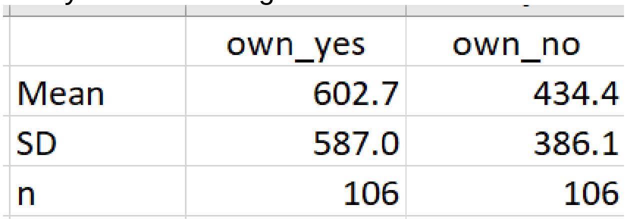
```
untidy_bookings <- Hotel_bookings %>%
        group_by(hotel) %>%
        summarise (cancel_proportion = 100 * (sum(is_cancelled == 1) / n())) %>%
        mutate (not_cancel_proportion = 1 – cancel_proportion)

tidy_bookings <- untidy_bookings %>%
        pivot_longer (cols = c(cancel_proportion, not_cancel_proportion), name_to =
"Cancellation_Status", values_to = "proportion")

ggplot (tidy_bookings, mapping = aes (x = hotel, y = proportion, fill = Cancellation_Status)) +
geom_col (position = "dodge")+
labs (x = "Hotel Type", y = "", title = "Cancellation Status % by Hotel Type")
```

Question 2 (20 points)

A bank wants to study whether there is any difference in credit card balances between those who own their house (***own_yes***) and those who don't (***own_no***). They collected a sample of customers and the summary statistics are given below:



	own_yes	own_no
Mean	602.7	434.4
SD	587.0	386.1
n	106	106

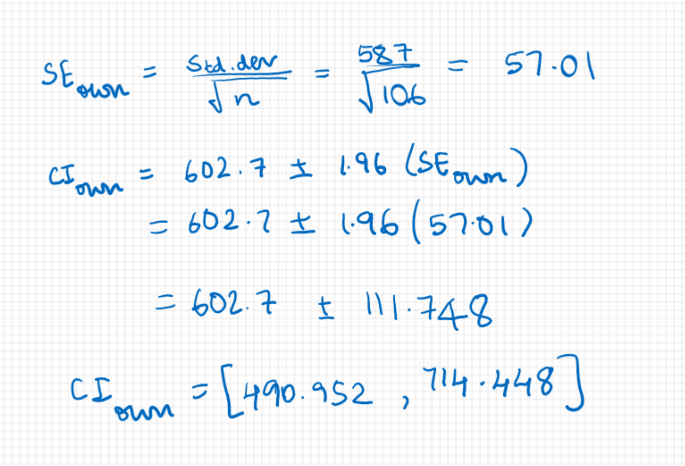
Part a (3 points): Please state what is the population, the sample, the parameter you want to infer, and the available sample statistic

- The population represents all the clients of the bank.
- The sample is 212 clients (106 people who own a house, 106 people who do not own a house) who participated in the survey.
- Parameter is the difference of mean of credit card balances of the people who own and do not own a house.
- Sample statistics = 602.7 – 434.4 = 168.3

Part b (7 points) Construct two 95% confidence intervals; one for the average card balance for those who own their house (own_yes) and one for those who don't (own_no). Do you have to make any assumptions?

- For the people who own their own house
  - SE(Standard Error) = 57.01
  - Confidence Intervals = 602.7 +- (1.96*SE)

    = [490.952 , 714.448]



$$SE_{own} = \frac{Std.dev}{\sqrt{n}} = \frac{587}{\sqrt{106}} = 57.01$$

$$CI_{own} = 602.7 \pm 1.96\,(SE_{own})$$

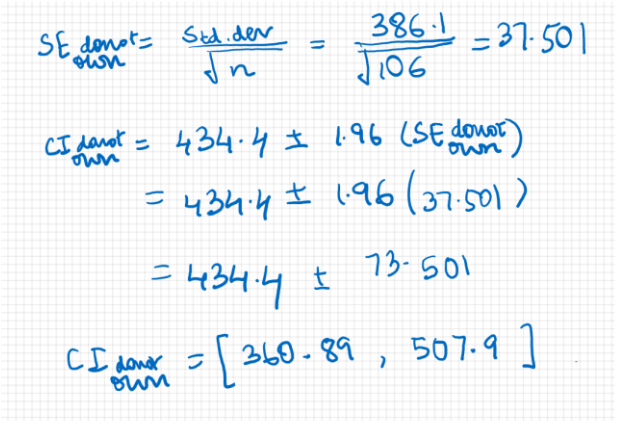
$$= 602.7 \pm 1.96\,(57.01)$$

$$= 602.7 \pm 111.748$$

$$CI_{own} = [490.952\ ,\ 714.448]$$

- For the people who do not own their own house
    - SE(Standard Error) = 37.501
    - Confidence Intervals = 434.4 +- (1.96*SE)

      = [360.89 , 507.9]



$$SE_{\substack{donot\\own}} = \frac{Std.dev}{\sqrt{n}} = \frac{386.1}{\sqrt{106}} = 37.501$$

$$CI_{\substack{donot\\own}} = 434.4 \pm 1.96\ (SE_{\substack{donot\\own}})$$

$$= 434.4 \pm 1.96\left(37.501\right)$$

$$= 434.4 \pm 73.501$$

$$CI_{\substack{donot\\own}} = \left[360.89\ ,\ 507.9\right]$$

Assumption - No assumptions are needed. There is sufficient sample size n>30 in this example, hence we can say that the sample mean follows a Normal Distribution according to Central Limit Theorem.

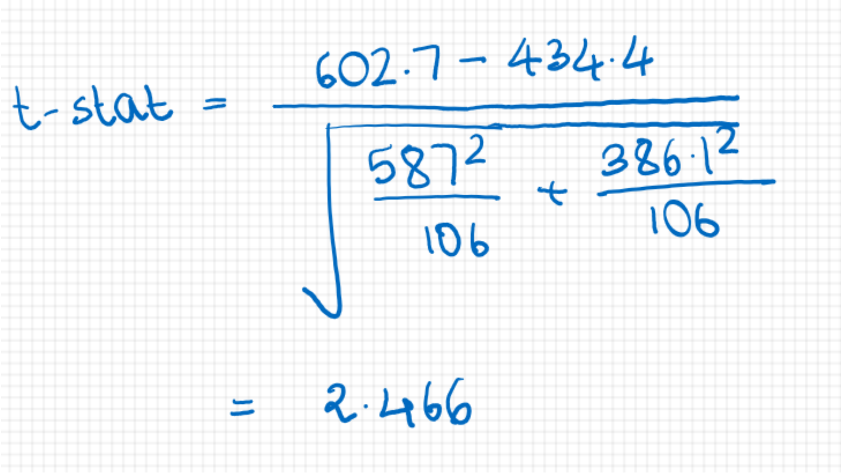
Part c (10 points) Based on this sample, test whether or not the mean difference of credit card balances for those customers who own their house is the same as those who don't. Use a 5% significance level. Conduct a hypothesis testing, state the null and the alternative, calculate a t-statistic for the difference, and finally state what you decide/infer.

- NULL HYPOTHESIS (H_0)- There is no difference in the credit card balances of people who own a house and those who do not own a house.

  i.e. difference (mu_own_house - mu_do_not_own_house) = 0

- ALTERNATIVE HYPOTHESIS (H_a)- There is a difference in the credit card balances of people who own a house and those who do not own a house.

  i.e. difference (mu_own_house - mu_do_not_own_house) not equal to 0
- Calculations



$$\text{t-stat} = \frac{602.7 - 434.4}{\sqrt{\frac{587^2}{106} + \frac{386.1^2}{106}}}$$

$$= 2.466$$

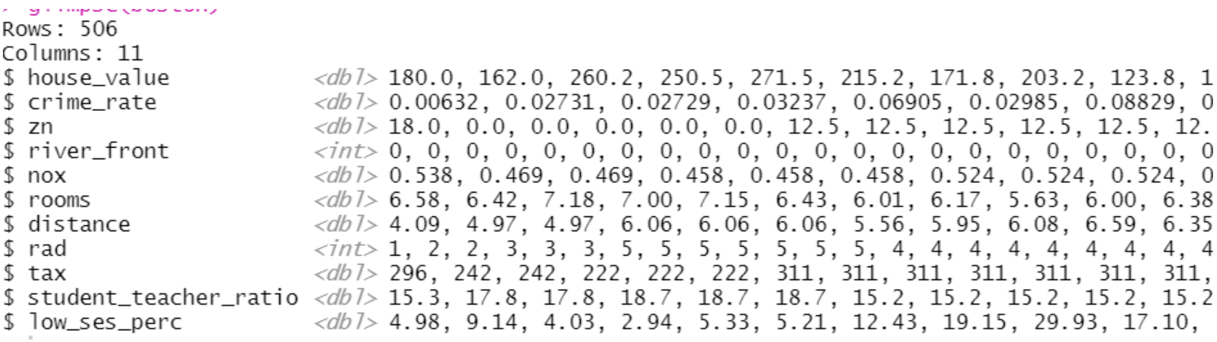
With confidence level of 95%, we can reject the null hypothesis since T-statistic is greater than T critical (approximately 2)
(T stats > T critical, the p-value should be smaller than 5%. Hence, reject H_0!)

There is a significant difference in mean of credit card balances of people who own a house and those who do not own a house.

T stats > T critical, the p-value should be smaller than 5%. Conclusion: reject H_0

Question 3 (25 points)

The following data is about median house value (***house_value***, in thousands of $) for 506 census districts in the greater Boston are. The variables are as follows

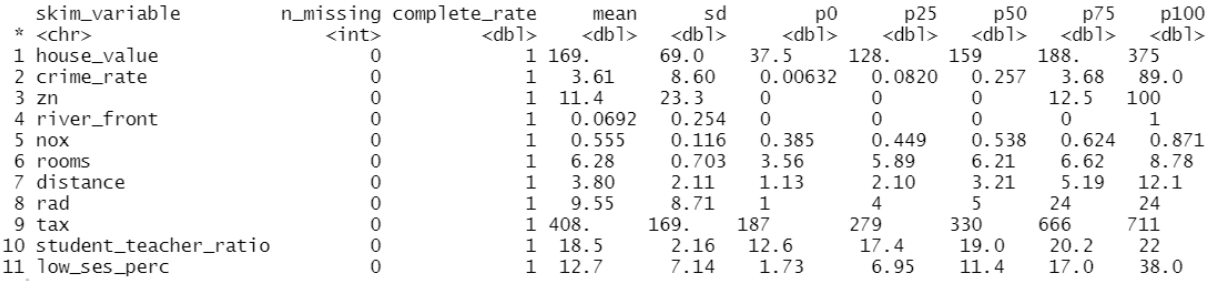


```
> glimpse(boston)
Rows: 506
Columns: 11
$ house_value          <dbl> 180.0, 162.0, 260.2, 250.5, 271.5, 215.2, 171.8, 203.2, 123.8, 1
$ crime_rate           <dbl> 0.00632, 0.02731, 0.02729, 0.03237, 0.06905, 0.02985, 0.08829, 0
$ zn                   <dbl> 18.0, 0.0, 0.0, 0.0, 0.0, 0.0, 12.5, 12.5, 12.5, 12.5, 12.5, 12.
$ river_front          <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
$ nox                  <dbl> 0.538, 0.469, 0.469, 0.458, 0.458, 0.458, 0.524, 0.524, 0.524, 0
$ rooms                <dbl> 6.58, 6.42, 7.18, 7.00, 7.15, 6.43, 6.01, 6.17, 5.63, 6.00, 6.38
$ distance             <dbl> 4.09, 4.97, 4.97, 6.06, 6.06, 6.06, 5.56, 5.95, 6.08, 6.59, 6.35
$ rad                  <int> 1, 2, 2, 3, 3, 3, 5, 5, 5, 5, 5, 5, 5, 4, 4, 4, 4, 4, 4, 4, 4, 4
$ tax                  <dbl> 296, 242, 242, 222, 222, 222, 311, 311, 311, 311, 311, 311, 311,
$ student_teacher_ratio <dbl> 15.3, 17.8, 17.8, 18.7, 18.7, 18.7, 15.2, 15.2, 15.2, 15.2, 15.2
$ low_ses_perc         <dbl> 4.98, 9.14, 4.03, 2.94, 5.33, 5.21, 12.43, 19.15, 29.93, 17.10,
```

crime_rate: per capita crime rate by town.
zn: proportion of residential land zoned for lots over 25,000 sq.ft.
river_front: Charles River dummy variable (= 1 if tract bounds river; 0 otherwise).
nox: nitrogen oxides concentration (parts per 10 million).
rooms: average number of rooms per dwelling.
age: proportion of owner-occupied units built prior to 1940.
distance: weighted mean of distances to five Boston employment centres.
rad: index of accessibility to radial highways.
tax: full-value property-tax rate per $10,000.
student_teacher_ratio: Student-teacher ratio by town.
low_ses_perc: lower status of the population (percent).

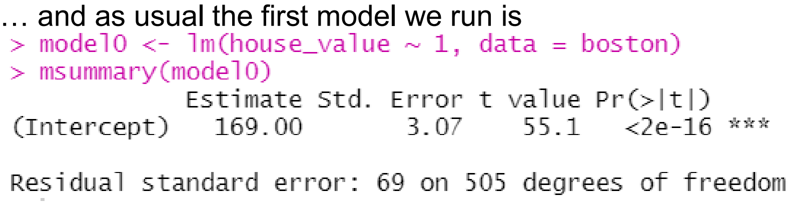
Part a (5 pts)

The following output summarises the data set



```
   skim_variable         n_missing complete_rate    mean      sd        p0      p25     p50     p75    p100
 * <chr>                     <int>         <dbl>   <dbl>   <dbl>     <dbl>    <dbl>   <dbl>   <dbl>   <dbl>
 1 house_value                   0             1 169.     69.0    37.5     128.     159     188.    375
 2 crime_rate                    0             1   3.61    8.60    0.00632   0.0820   0.257   3.68   89.0
 3 zn                            0             1  11.4    23.3     0         0        0      12.5   100
 4 river_front                   0             1   0.0692  0.254   0         0        0       0       1
 5 nox                           0             1   0.555   0.116   0.385     0.449    0.538   0.624   0.871
 6 rooms                         0             1   6.28    0.703   3.56      5.89     6.21    6.62    8.78
 7 distance                      0             1   3.80    2.11    1.13      2.10     3.21    5.19   12.1
 8 rad                           0             1   9.55    8.71    1         4        5      24      24
 9 tax                           0             1 408.    169.    187       279      330     666     711
10 student_teacher_ratio         0             1  18.5     2.16   12.6      17.4     19.0    20.2    22
11 low_ses_perc                  0             1  12.7     7.14    1.73      6.95    11.4    17.0    38.0
```

… and as usual the first model we run is



```
> model0 <- lm(house_value ~ 1, data = boston)
> msummary(model0)
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   169.00       3.07    55.1   <2e-16 ***

Residual standard error: 69 on 505 degrees of freedom
```

Without looking at a graph, how would you determine whether the distribution of ***house_value*** is symmetric or skewed? What is a 95% confidence interval for the average of ***house_value***?

- House_value is a Right Skewed distribution – this is because the median (159) is less than the mean (169)

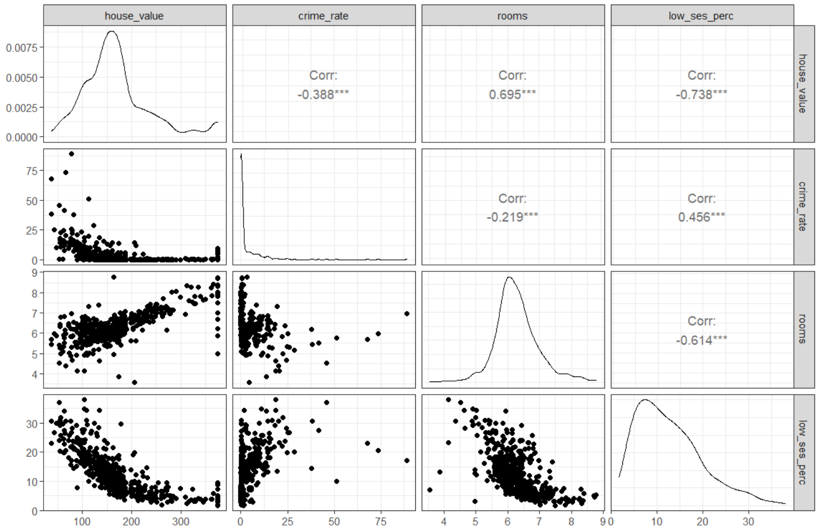
- house_value statistics:
  mean = 169
  standard deviation = 69
  n = 506 (and there are no missing values)

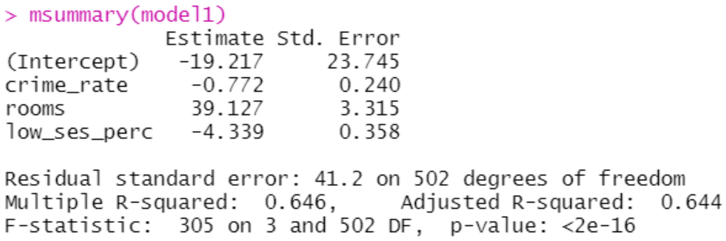
$$\text{Std. error} = \frac{69}{\sqrt{506}} = 3.07$$

$$\begin{aligned} CI &= 169 \pm (1.96 * 3.07) \\ &= 169 \pm 6.0172 \\ &= [162.98,\ 175.017] \end{aligned}$$

Part b (5 pts)

You use ***crime_rate***, ***rooms***, and ***low_ses_perc*** to fit your first model to explain ***house_value***. Below is the scatterplot matrix and the regression output.



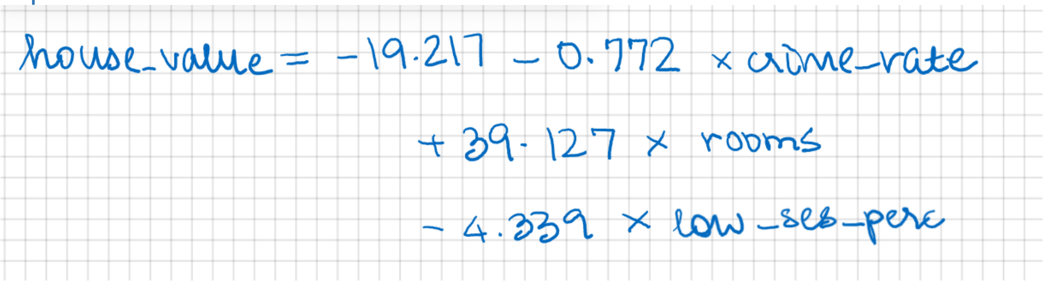


```
> msummary(model1)
             Estimate Std. Error
(Intercept)   -19.217     23.745
crime_rate     -0.772      0.240
rooms          39.127      3.315
low_ses_perc   -4.339      0.358

Residual standard error: 41.2 on 502 degrees of freedom
Multiple R-squared:  0.646,     Adjusted R-squared:  0.644
F-statistic:  305 on 3 and 502 DF,  p-value: <2e-16
```

Write out the equation of the linear model and escribe the results obtained (2pt).
Are the slopes you estimated significant? (2pts)
What proportion of the variability in *house_value* does your model explain? (1pts)

Equation of the linear model



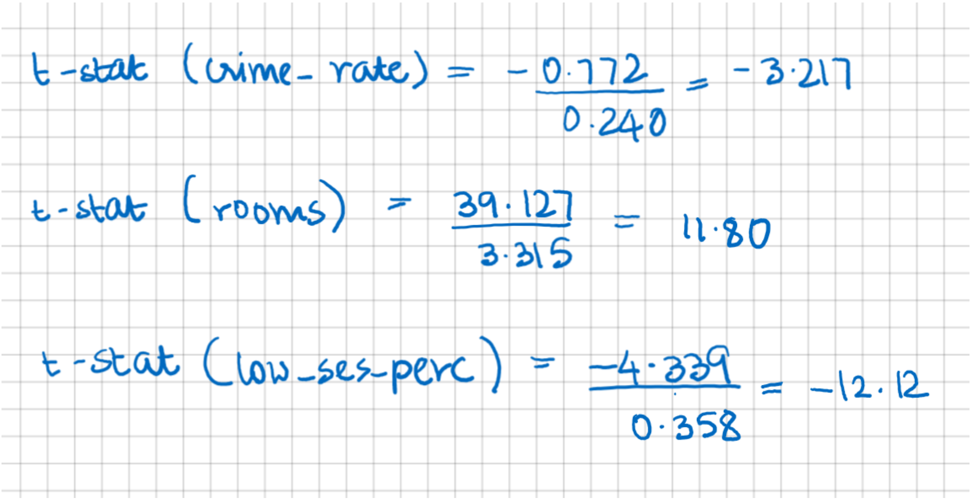
$$\text{house\_value} = -19.217 - 0.772 \times \text{crime\_rate} + 39.127 \times \text{rooms} - 4.339 \times \text{low\_ses\_perc}$$

Description of Results

- The intercept is -19.217 which can be interpreted as when the crime rate, rooms and lower status of the population (percent) are 0, then the median house value of that district is -19.217 which is not a realistic situation
- The crime rate coefficient is -0.772 which means that with a 1% increase in crime rate, the median house value of that district goes down by 0.772 times thousand which is $772
- The coefficient of the room is 39.127 which means the when the average number of rooms per house increases by 1, the median house value of that district increases by approximately $40,000
- If the percentage of people in the lower status increases by 1%, the median house value of that district decreases by around $4400.

Are the slopes significant?

- Calculating the t-statistic values for the coefficients to check which ones are significant:



$$\text{t-stat (crime-rate)} = \frac{-0.772}{0.240} = -3.217$$

$$\text{t-stat (rooms)} = \frac{39.127}{3.315} = 11.80$$

$$\text{t-stat (low\_ses\_perc)} = \frac{-4.339}{0.358} = -12.12$$

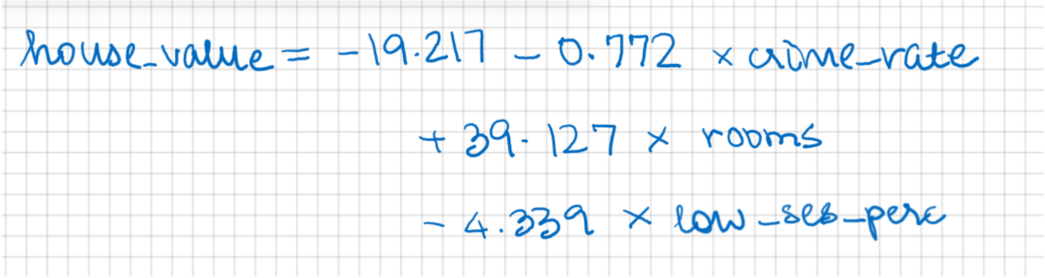
From the calculations we can see that the absolute t-value of crime rate, rooms and lower status population are all more than 2, and hence all the slopes are significant.

The model explain 64.4% of the variability in house_value

Part c (5 pts)

Using model 1, please construct a 95% prediction interval for ***house_value,*** given that crime_rate = 4, rooms = 6, and low_ses_perc = 15.

Equation used-



$$\text{house\_value} = -19.217 - 0.772 \times \text{crime\_rate} + 39.127 \times \text{rooms} - 4.339 \times \text{low\_ses\_perc}$$

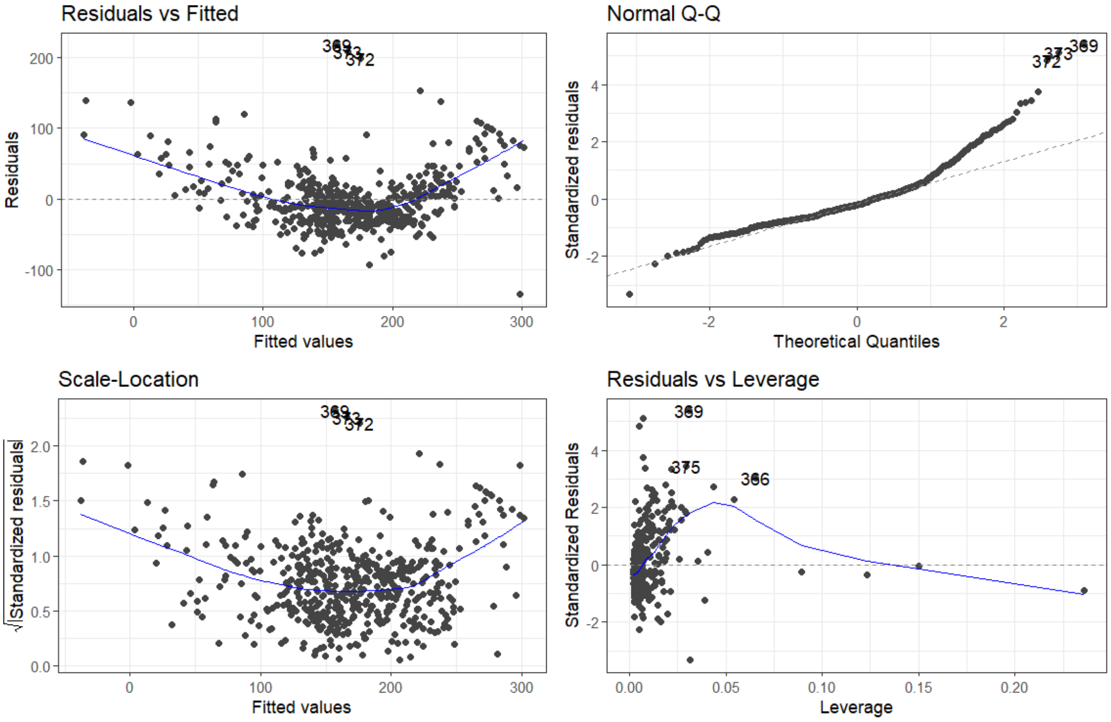
House_value = -19.217- (0.772*crime_rate) + (39.127*rooms)
– (4.339*low_ses_perc)
= -19.217 – (0.772*4) + (39.127*6) – (4.339*15)
= 147.372

Standard error = Residual standard error of model = 41.2

Constructing 95% prediction interval for house_value-
CI = 147.372 +- (1.96* 41.2)
[66.62, 228.124]

Part d (5 pts)

The following are the diagnostic plots for model 1. What do you see? How could you improve model 1?



Residuals vs. Fitted: there is a pattern in the data that is currently unaccounted for. The data is following a sort of quadratic trend.

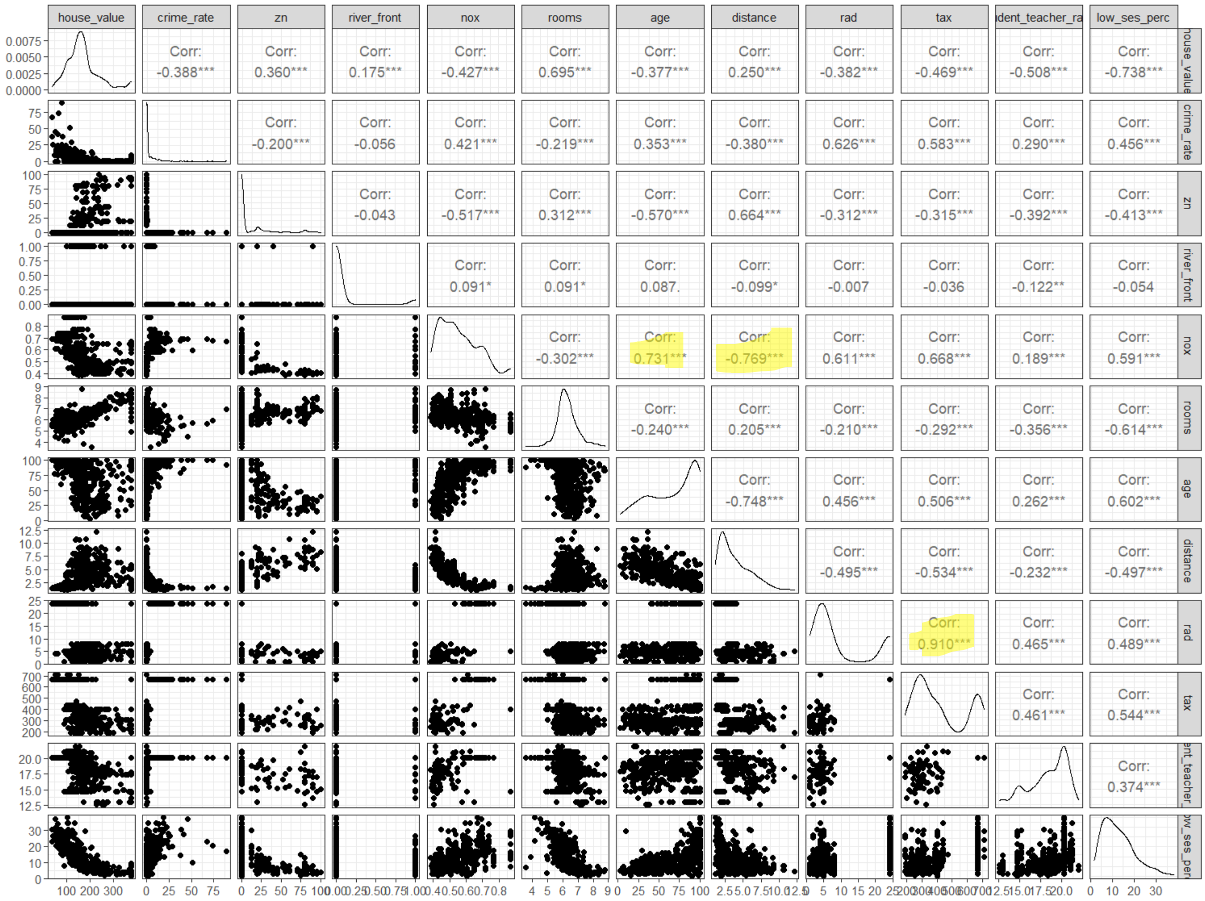
Normal Q-Q: Residuals are not following a normal distribution pattern as per the graph shown

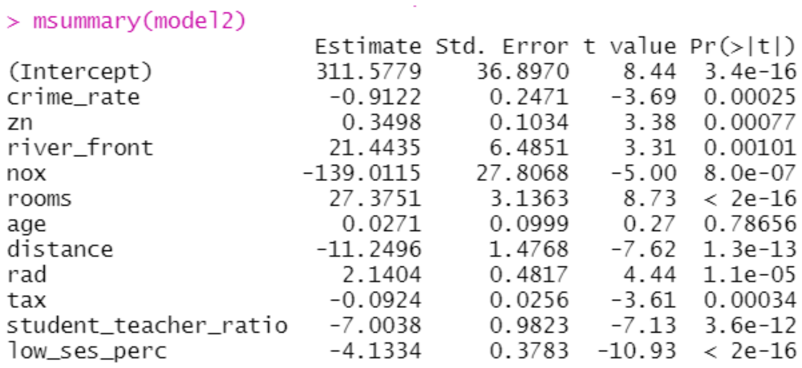
Scale-Location: Positive or negative trends across the fitted values indicate variability that is not constant.

Residuals vs. Leverage:

Points with high leverage (having unusual values of the predictors) and/or high absolute residuals can have an undue influence on estimates of model parameters.

Part e (5 pts) You now decide to fit a model with all predictors; the scatterplot matrix and associated regression output is given below





```
> msummary(model2)
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            311.5779    36.8970    8.44  3.4e-16
crime_rate              -0.9122     0.2471   -3.69  0.00025
zn                       0.3498     0.1034    3.38  0.00077
river_front             21.4435     6.4851    3.31  0.00101
nox                   -139.0115    27.8068   -5.00  8.0e-07
rooms                   27.3751     3.1363    8.73  < 2e-16
age                      0.0271     0.0999    0.27  0.78656
distance               -11.2496     1.4768   -7.62  1.3e-13
rad                      2.1404     0.4817    4.44  1.1e-05
tax                     -0.0924     0.0256   -3.61  0.00034
student_teacher_ratio   -7.0038     0.9823   -7.13  3.6e-12
low_ses_perc            -4.1334     0.3783  -10.93  < 2e-16
```

Explain the effect of ***river_front.*** Is it significant? [2 pts]
Is there anything that would make you worried about model 2? [3 pts]

river_front is a significant variable since the t-value is 3.31 which is more than 2.

river_front variable can be interpreted as:
Every house group in the district that is bounded by the river has a median house value almost $21,443 more than the districts that are not bound by the river.

Concerns in the model-

- The variable representing age i.e. the proportion of owner-occupied units built prior to 1940 is not significant since it has a t-value of 0.27 (which is <2) and hence it can be omitted from the model
- From the list of scatter plot matrix shown, it can be observed that several variables have a correlation of more than 60%, but for now we can prioritize looking into the following variables:
    - i. Both “nox” and “distance” are used in the model as explanatory variables but the two variables have a correlation of 76.9%
    - ii. Both “nox” and “age” are used in the model as explanatory variables but the two variables have a correlation of 73.1%
    - iii. Both “tax” and “rad” are used in the model as explanatory variables but the two variables have a correlation of 91%
- As next steps, we can calculate the Variance Inflation Factors (vif) for the explanatory variables used in the model, drop the variables with high correlation and re-run regression.