Week 4 Tasks

1 Tasks

You are encouraged to complete the following tasks by using Quarto to produce a single document which summarises all your work, i.e. the original questions, your R code, your comments and reflections, etc.

1.1 Part 1

Data was collected on the characteristics of homes in the American city of Los Angeles (LA) in 2010 and can be downloaded below:

The data contain the following variables:

- `city` - the district of LA where the house was located
- `type` - either `SFR` (Single Family Residences) or `Condo/Twh` (Condominium/Town House)
- `bed` - the number of bedrooms
- `bath` - the number of bathrooms
- `garage` - the number of car spaces in the garage
- `sqft` - the floor area of the house (in square feet)
- `pool` - `Y` if the house has a pool
- `spa` - `TRUE` if the house has a spa
- `price` - the most recent sales price ($US)

We are interested in exploring the relationships between `price` and the other variables.

Read the data into an object called `LAhomes` and complete the following tasks

```
LAhomes <- read.csv("LAhomes.csv",stringsAsFactors = T)
```

Task 1

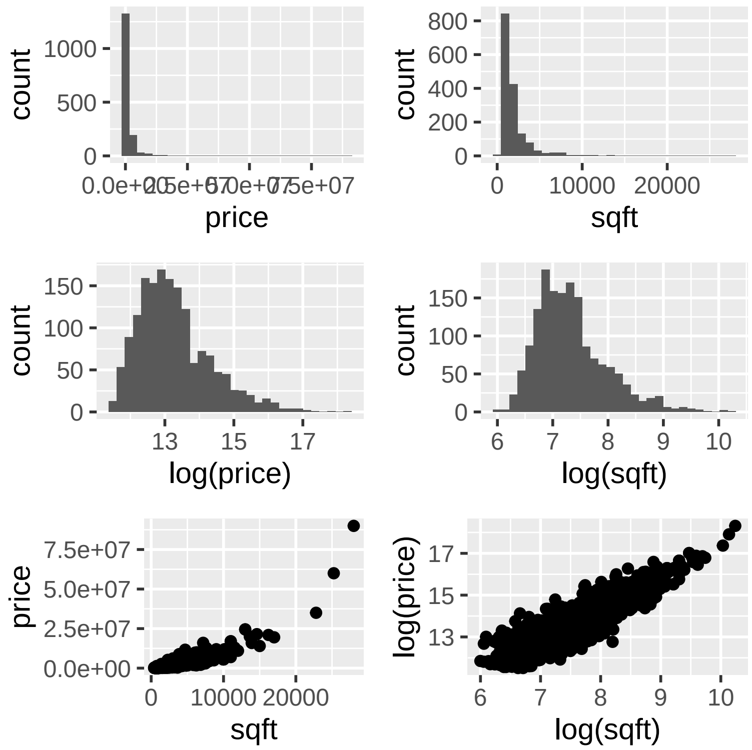
By looking at the univariate and bivariate distributions on the `price` and `sqft` variables below, what would be a sensible way to proceed if we wanted to model this data? What care must be taken if you were to proceed this way?
Click here to see the solution

```r
hist1 <- ggplot(LAhomes, aes(x = price)) +
          geom_histogram()
hist2 <- ggplot(LAhomes, aes(x = sqft)) +
          geom_histogram()

# Explore log transformation
hist1log <- ggplot(LAhomes, aes(x = log(price))) +
             geom_histogram()
hist2log <- ggplot(LAhomes, aes(x = log(sqft))) +
             geom_histogram()

plot1 <- ggplot(LAhomes, aes(x = sqft, y = price)) +
          geom_point()
plot2 <- ggplot(LAhomes, aes(x = log(sqft), y = log(price))) +
          geom_point()

grid.arrange(hist1, hist2, hist1log, hist2log, plot1, plot2,
             ncol = 2, nrow = 3)
```

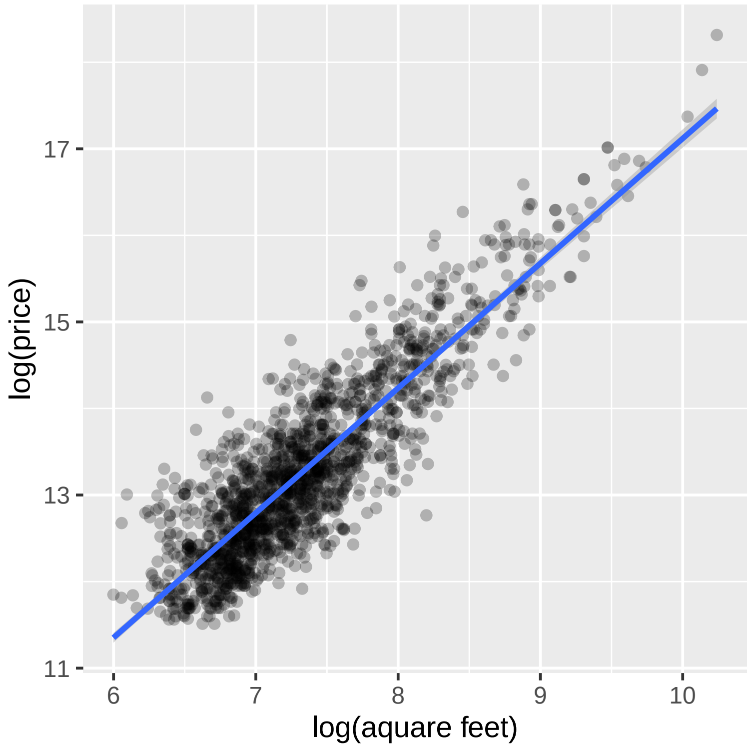


Task 2

Fit the simple linear model with `log(price)` as the response and `log(sqft)` as the predictor. Display the fitted model on a scatterplot of the data and construct a confidence interval for the slope parameter in the model and interpret its point and interval estimates.
Click here to see the solution

```r
slr_LAprices <- lm(log(price) ~ log(sqft), data = LAhomes)

ggplot(LAhomes, aes(x =  log(sqft), y = log(price))) +
  geom_point(alpha=0.25) +
  labs(x = "log(aquare feet)", y = "log(price)") +
  geom_smooth(method = "lm", level =0.95)
```




```r
tab_model(slr_LAprices)
```

		log(price)	
Predictors	Estimates	CI	p
(Intercept)	2.70	2.42 – 2.98	**<0.001**
sqft [log]	1.44	1.40 – 1.48	**<0.001**
Observations	1594		
R^2 / R^2 adjusted	0.774 / 0.774		

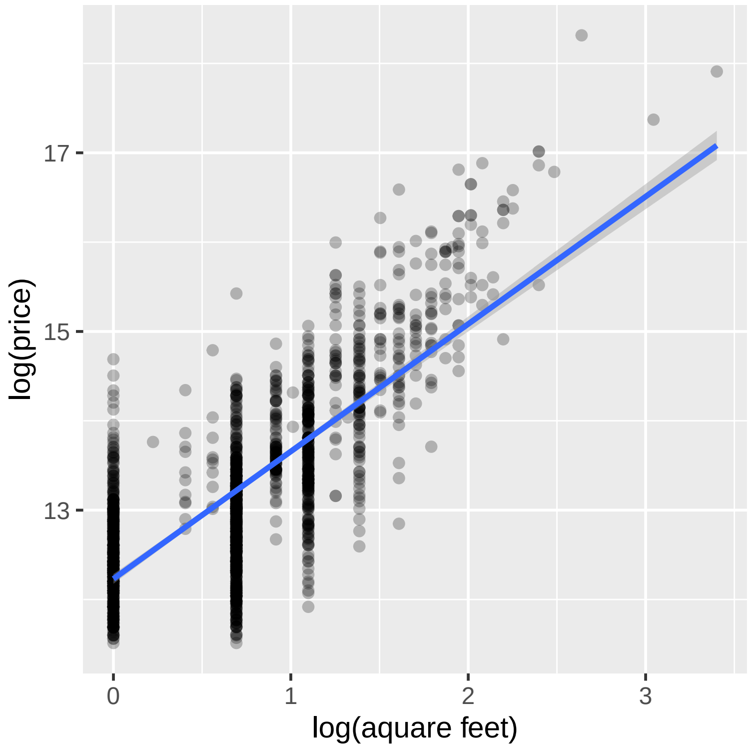
Task 3

Re-do your analysis but now using `log(bath)` as the explanatory variable. Calculate the point and interval estimates of the coefficients.
Click here to see the solution

```r
slr_LAprices2 <- lm(log(price) ~ log(bath), data = LAhomes)

ggplot(LAhomes, aes(x =  log(bath), y = log(price))) +
  geom_point(alpha=0.25) +
  labs(x = "log(aquare feet)", y = "log(price)") +
  geom_smooth(method = "lm", level =0.95)
```

```
`geom_smooth()` using formula = 'y ~ x'
```




```r
tab_model(slr_LAprices2)
```

		log(price)	
Predictors	Estimates	CI	p
(Intercept)	12.23	12.18 – 12.29	**<0.001**
bath [log]	1.43	1.37 – 1.49	**<0.001**
Observations	1594		
R^2 / R^2 adjusted	0.577 / 0.577		

Task 4

Fit the multiple linear regression model using the **log transform of all the variables** `price` (as the response) and both `sqft` and `bath` (as the explanatory variables). Calculate the point and interval estimates of the coefficients of the two predictors separately. Compare their point and interval estimates to those you calculated before. Can you account for the differences?
Click here to see the solution

```r
mlr_LAprices <- lm(log(price) ~ log(sqft) + log(bath), data = LAhomes)
tab_model(mlr_LAprices,slr_LAprices,slr_LAprices2)
```

	log(price)			log(price)			log(price)	
Predictors	Estimates	CI	p	Estimates	CI	p	Estimates	CI
(Intercept)	2.51	2.00 – 3.03	**<0.001**	2.70	2.42 – 2.98	**<0.001**	12.23	12.18 – 12.29
sqft [log]	1.47	1.39 – 1.55	**<0.001**	1.44	1.40 – 1.48	**<0.001**		
bath [log]	-0.04	-0.13 – 0.05	0.389				1.43	1.37 – 1.49
Observations	1594			1594			1594	
R^2 / R^2 adjusted	0.774 / 0.774			0.774 / 0.774			0.577 / 0.577	

Task 5

Using the objective measures for model comparisons, which of the models in task 2 ,3 and 4 would you favour? Is this consistent with your conclusions in task 4?
Click here to see the solution

```r
tab_model(mlr_LAprices,slr_LAprices,slr_LAprices2,show.aic = T)
```

	log(price)			log(price)			log(price)	
Predictors	Estimates	CI	p	Estimates	CI	p	Estimates	CI
(Intercept)	2.51	2.00 – 3.03	**<0.001**	2.70	2.42 – 2.98	**<0.001**	12.23	12.18 – 12.29
sqft [log]	1.47	1.39 – 1.55	**<0.001**	1.44	1.40 – 1.48	**<0.001**		
bath [log]	-0.04	-0.13 – 0.05	0.389				1.43	1.37 – 1.49
Observations	1594			1594			1594	
R^2 / R^2 adjusted	0.774 / 0.774			0.774 / 0.774			0.577 / 0.577	
AIC	44587.722			44586.467			45584.113	

1.2 Part 2.

You have been asked to determine the pricing of a New York City (NYC) Italian restaurant's dinner menu such that it is competitively positioned with other high-end Italian restaurants by analysing pricing data that have been collected in order to produce a regression model to predict the price of dinner.

Data from surveys of customers of 168 Italian restaurants in the target area are available. The data can be found in the file `restNYC.csv`.

Each row represents one customer survey from Italian restaurants in NYC and includes the key variables:

- `Price` - price (in $US) of dinner (including a tip and one drink)
- `Food` - customer rating of the food (from 1 to 30)
- `Decor` - customer rating of the decor (from 1 to 30)
- `Service` - customer rating of the service (from 1 to 30)
- `East` - dummy variable with the value 1 if the restaurant is east of Fifth Avenue, 0 otherwise

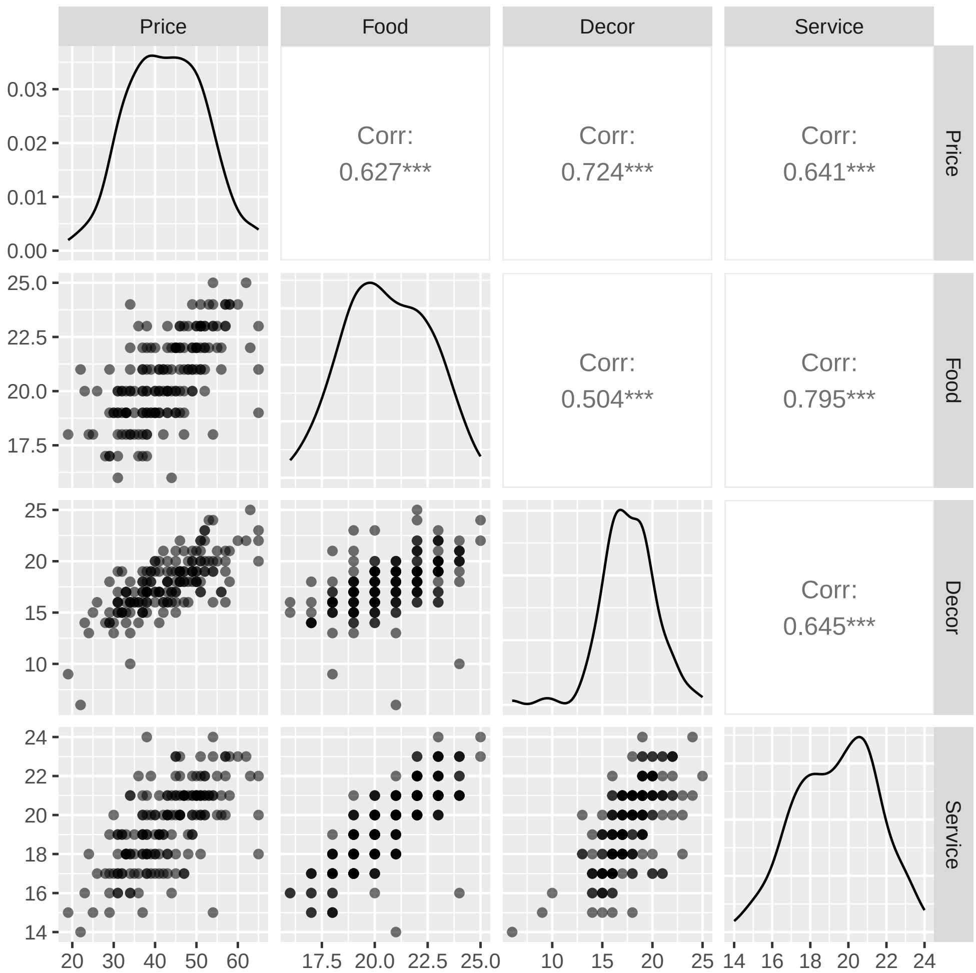
```r
restNYC <- read.csv("restNYC.csv",stringsAsFactors = T)
```

Task 6

Using the ggpairs function in the GGally package (see the following code) we can generate an informative set of graphical and numerical summaries which illuminate the relationships between pairs of variables. Where do you see the strongest evidence of relationships between price and the potential explanatory variables? Is there evidence of multicollineatity in the data?

```
library(GGally) # Package to produce matrix of 'pairs' plots and more!
restNYC$East <- as.factor(restNYC$East) # East needs to be a factor
# Including the `East` factor
# ggpairs(restNYC[, 4:8], aes(fill = East, alpha = 0.4),progress = F)

# Without the `East` factor
ggpairs(restNYC[, 4:7], aes(alpha = 0.4),progress = F)
```

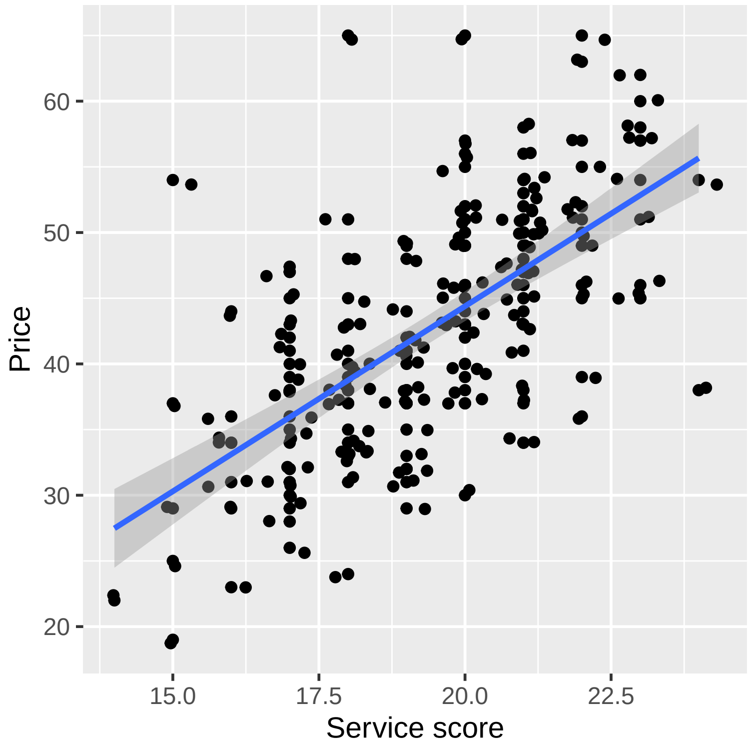


Solution
There seems to be a strong positive linear association between service and food, which could lead to some collinearity issues.

Task 7

Fit the simple linear model with `Price` as the response and `Service` as the predictor and display the fitted model on a scatterplot of the data. Construct a confidence interval for the slope parameter in the model.
Click here to see the solution

```
price_serv_LM <- lm(Price  ~ Service  , data = restNYC)
ggplot(restNYC, aes(x = Service, y = Price)) +
  geom_point() +
  geom_jitter() +
  labs(x = "Service score", y = "Price")+
  geom_smooth(method = "lm")
```




```
broom::tidy(price_serv_LM,conf.int = T)
```

```
# A tibble: 2 x 7
  term        estimate std.error statistic  p.value conf.low conf.high
  <chr>          <dbl>     <dbl>     <dbl>    <dbl>    <dbl>     <dbl>
1 (Intercept)   -12.0      5.11      -2.34 2.02e- 2   -22.1      -1.89
2 Service         2.82     0.262     10.8  7.88e-21     2.30      3.34
```

Task 8

Now fit a multiple regressing model of `Price` on `Service`, `Food`, and `Decor`. What happens to the significance of `Service` when additional variables were added to the model?
Click here to see the solution

```r
price_serv_MLR <- lm(Price  ~ Service + Food + Decor , data = restNYC)
tab_model(price_serv_MLR,collapse.ci = T)
```

	Price	
Predictors	Estimates	p
(Intercept)	-24.64 (-34.03 – -15.25)	**<0.001**
Service	0.14 (-0.65 – 0.92)	0.733
Food	1.56 (0.82 – 2.29)	**<0.001**
Decor	1.85 (1.42 – 2.28)	**<0.001**
Observations	168	
R^2 / R^2 adjusted	0.617 / 0.610	

Task 9

What is the correct interpretation of the coefficient on `Service` in the linear model which regresses `Price` on `Service`, `Food`, and `Decor`?

See solution

After controlling for `Food` and `Decor`, a 1-point increase in the `Service` rating is associated with an estimated $0.135 increase in the. average Price, but this effect is not statistically significant ($p > 0.05$)