1. **[7 Total Marks]** Aaron Judge, a baseball player playing for the New York Yankees, just hit the most Home Run (or HR) in the Major League Baseball (or MLB) this regular season. He has played 9 seasons in the MLB and here are the number of Home Runs he hit from the 2016 season to 2024 season. Note that the 2020 season was shortened due to the COVID pandemic.

Season	2016	2017	2018	2019	2020*	2021	2022	2023	2024
# of HR	4	52	27	27	9	39	62	37	58

a) Find the median of the number of home run per season. **[2 marks]**

# of HR 4 9 27 27 37 39 52 58 62

$i = \frac{n+1}{2} = \frac{9+1}{2} = 5$

Median = 37

2

b) Use the method in our notes to find the IQR. **[3 marks]**

$Q_1 = 18$

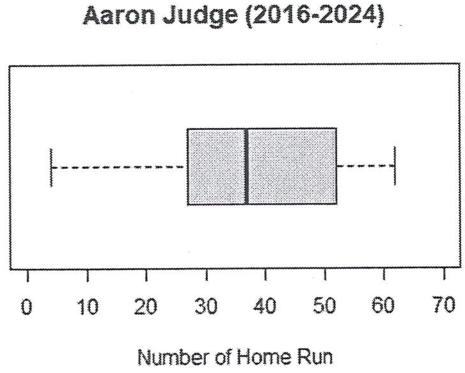
$Q_3 = 55$

$IQR = Q_3 - Q_1 = 55 - 18 = 37$

$IQR = 37$

3

c) Provide a description of the boxplot below. **[2 marks]**
Note: Don't worry about the discrepancy of Q1 and Q3 values.



2

The median is aprox. 37
The $Q_1 \simeq 27$ and $Q_2 \simeq 52$
the $IQR \simeq 25$
The graph is right-skewed and there are not any visible outliers.

2. **[8 Total Marks]** A local news reporter wants to find out if October is too early to put up Christmas decorations (like Christmas trees and outdoor lightings etc.) in the City of Vancouver. She divides the City of Vancouver into 200 neighbourhood blocks and 10 neighbourhood blocks are randomly chosen. She will then ask the resident(s) in every house (or unit) in those 10 neighbourhood blocks if they think putting up Christmas decorations in October is too early.

   a) Identify the subjects of interest. **[2 marks]**

   the residents in every house in the City of Vancouver in October.

   2

   b) Provide a description of the variable of interest. **[1 mark]**

   to find out if October is too early to put up Christmas decorations

   1

   c) Based on your answer in the previous part, identify the type of the variable. **[1 mark]**

   Categorical variable, class Yes/No answer.

   1

   d) Identify the most appropriate sampling method adopted by the reporter. Briefly justify your choice. **[1+3 marks]**

   Cluster sampling Method

   Because she divided the City of Vancouver into 200 neighbourhood blocks (clusters) and then she selected 10 clusters.

~~1.5~~

2.5

3. **[9 Total Marks]** In each of the following situations, (1) identify the most appropriate graph and (2) identify the most appropriate statistic to summarize the collected data, and (3) briefly provide a justification of the answer.

a) A researcher wanted to find out if the primary source of news (newspaper, TV/radio, and social media) varies among different age groups (under 36, 36-50, over 50). A random sample of 500 respondents was obtained for this research. **[3 marks]**

3

1) Side-by-side bar graph
2) X^2-statistic
3) They are two categorical variables and the researcher wanted to find out if they are independent to each other.

b) A medical school instructor wants to see how the length of time students spend working in a hospital during their residency (the number of total hours) is related to their board exam performance. The instructor randomly selects 42 of her students for this investigation. She asks them how many hours they have worked in the hospital during the last year of their residency, and also records the number of points they scored on the final board exam. **[3 marks]**

1) Scatterplot graph
2) covariance and correlation coefficient (r)
3) they are two numerical variables and the medical school instructor wants to see their correlation.

3

c) In order to estimate the average household income in Vancouver, an urban researcher surveyed a random sample 300 households. Each household was asked to provide the household's total annual income (in dollars) in 2023. Note that income data is always right-skewed. **[3 marks]**

1) Histogram graph
2) Median, Q_1 and Q_3 and IQR
3) Because it is easy to see the right-skewed in a histogram graph and because it may have outliers it's better to use Median, if we don't see outliers we can use the average, the min and max and the SD.

2

4. **[7 Total Marks]** A researcher in the labour market wanted to see if the education background of wage-earning workers is independent to the type of work. From a random sample of 450 wage-earning workers, their education background (**Education**: bachelor's degree or higher vs. no) and the type of work (**Type**: Skilled positions vs. non-skilled positions) were collected. Partial results are summarized in the following two-way table.
Note: Skilled positions require workers to have special skills besides knowledge learned in the pre-job training.

		Education		
		Has bachelor or higher	Without bachelor	
Type	Skilled Positions	180	50	230
	Not Skilled Positions	120	100	220
		300	150	450

a) Explain the meaning of "independence" to the researcher, using the context. **[2 marks]**

In this context it's when both categorical variables don't affect each other, in this case the education background doesn't affect the type of work and viceversa.

in what ways?

1

b) Calculate the expected frequency of wage-earning workers without a bachelor's degree ~~or higher~~ and who are in a skilled position. Provide an interpretation of its value. **[1+1 marks]**

$$e = \frac{(150)(230)}{450} = 76.6$$

✓

✗

This value represents the number of wage-earning workers expected to find if the skilled positions were independent to the education background.

1.5

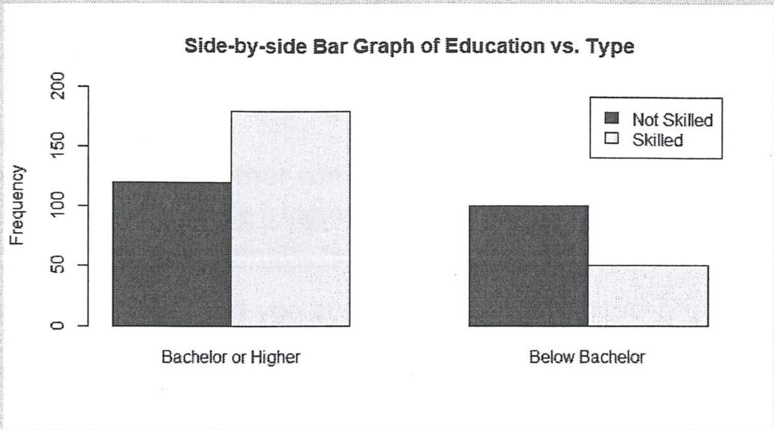
c) Provide the correct R code to make the side-by-side bar graph on the next page. **[3 marks]**
Note: You can use any of the R code provided in the two Reference sections.

Reference #1

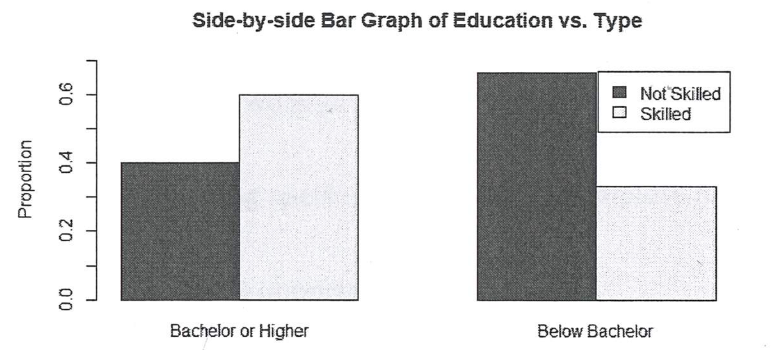
```
mydata <- data.frame(CatVar1, CatVar2)
mytab <- table(mydata)

# Find marginal totals - NULL=Table, 1=Row, 2=Column
margin.table(mytab) # Overall Total
margin.table(mytab, margin = 1) # Row Totals
margin.table(mytab, margin = 2) # Column Totals

# Find marginal % - NULL=Table, 1=Row, 2=Column
proportions(mytab)
proportions(mytab, margin = 1)
proportions(mytab, margin = 2)
```

Reference #2

```
mytab <- table(mydata)
barplot(mytab,
        beside = TRUE,
        legend.text = TRUE,
        main = "Side-by-side Bar Graph of Education vs. Type",
        ylab = "Frequency",
        ylim = c(0,200))
```



Answer



Your R code:

```
mydata <- data.frame(CatVar1, CatVar2)
mytab <- table (mydata)
proportions (mytab, margin=2)
barplot (proportions,
        beside = TRUE,
        legend.text = TRUE,
        main = "Side-by-side Bar Graph of Education vs. Type",
        ylab = "Proportion"
        ylim = c(0, 0.7))
```

3