Ishaan Taneja

Grading:
30 marks- Complete submission of Problem 1,3,5
70 marks- Problem 5

1. Suppose `x` is a vector. Describe what each of the below commands do.

```
> length(x)
> x[2]
> x[-2]
> x[1:5]
> x(length(x) -5 : length(x))
> x[c(1,3,5)]
> x[x>3]
> x[x<-2 | x>2]
> which(x == max(x))
```

Solution: 1

```
> length(x)
```

The command is used to get the length of vector x i.e. Number of elements in the vector x.

```
> x[2]
```

The command is used to extract the 2^{nd} element of vector x.

```
> x[-2]
```

The command is used to extract the all the elements of vector x except the 2^{nd} element.

```
> x[1:5]
```

The command is used to extract the first five elements of vector x.

```
> x[length(x) -5 : length(x)]
```

The command first generates the sequence from 5 to length of vector x, then each element of the sequence is subtracted from the length of vector x and the values at the respective resultant positions are extracted.
For example:

```
> x<-c(10,12,31,43,52,67,78,80)
> #length(x)=8
> #Sequence generated is 5,6,7,8 i.e. 5:length(x)
```

```
> #Resultant values are 8-5=3, 8-6=2, 8-7=1 and 8-8=0 i.e. length(x)- each value of sequen
> #So, the values at position 3,2 and 1 are extracted respectively.
> x[length(x) -5 : length(x)]

[1] 31 12 10

> x[c(1,3,5)]
```

The command is used to extract the 1^{st}, 3^{rd} and 5^{th} element of vector x.

```
> x[x>3]
```

The command is used to extract all the elements of vector x which are greater than 3.

```
> x[x< -2 | x>2]
```

The command is used to extract all the elements of vector x which are either greater than 2 or less than -2.

```
> which(x == max(x))
```

The command is used to extract the position(s) of greatest element of vector x.

2. Consider the dataset `diamonds` in `ggplot2` in R.

   (a) In two to three lines describing the dataset.

   (b) Write down the list of categories considered.

   (c) Construct a Bar Plot using the below command:

   i.
   ```
   > library(ggplot2)
   >  ggplot(data=diamonds) +
   +    geom_bar(mapping=aes(x=cut,  fill=clarity))+
   +      scale_fill_viridis_d()
   ```

   ii.
   ```
   > ggplot(data=diamonds) +
   +     geom_bar(mapping=aes(x=cut,  fill=clarity), position="dodge") +
   +     scale_fill_viridis_d()
   ```

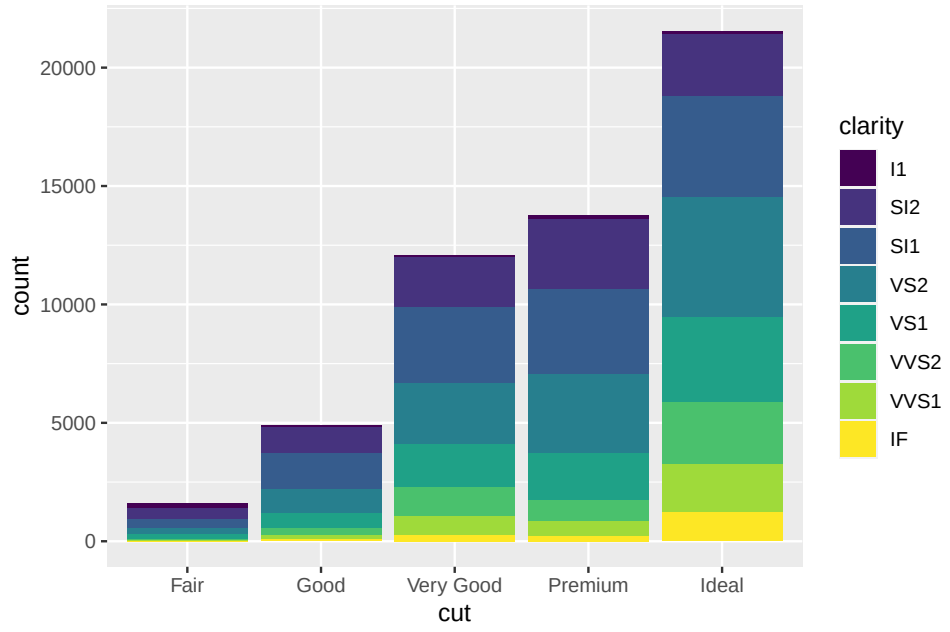
   and describe the differences in the outputs.

Solution: 2

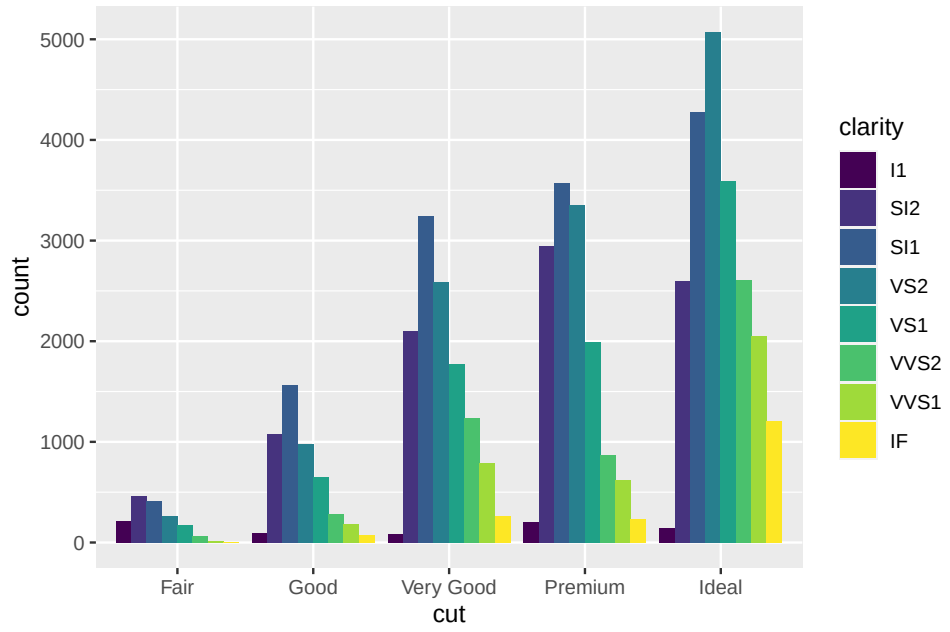
(a). The dataset `diamonds` in `ggplot2` in R, describes the price that vary between 326 USD and 18,823 USD of almost 54000 round cut diamonds.
It also describes other attributes of these diamonds like weights (vary between 0.2 and 5.01 carats), cut, color, clarity, dimensions, and more.

(b). The list of categories are price, carat, cut, clarity, color, length, width, depth, total depth percentage and width of top of diamond relative to its widest point.

(c). i.
```
> library(ggplot2)
>  ggplot(data=diamonds) +
+    geom_bar(mapping=aes(x=cut,  fill=clarity))+
+      scale_fill_viridis_d()
```



ii.
```
> ggplot(data=diamonds) +
+     geom_bar(mapping=aes(x=cut,  fill=clarity), position="dodge") +
+     scale_fill_viridis_d()
```



The command (i) produces a vertically stacked bar chart which gives the count for each category of cut quality of diamonds. For every category, each bar, in turn, shows the distribution of different clarify measures.
While command (ii) produces a grouped bar chart which represents the same thing

except that it is more convenient to look for the count on y axis for different clarity measures for each category of cut quality.

3. Load the package `UsingR` consider the dataset `cavendish`.

   (a) In two to three lines describing the dataset.

   (b) Provide the five number summary of the three variables considered using the `summary` function.

Solution: 3

(a). The dataset `Cavendish` consists of 29 observations and 3 variables which describes the mean density of earth. The 3 variables are 'density', 'density2' and 'density3' where, 'density2' is same as 'density' with the third value (4.88) replaced by 5.88 and 'density3' is also same as 'density' with first 6 observations ommitted.

(b).
```
> library(UsingR)
> summary(Cavendish)
```

```
    density         density2        density3
 Min.   :4.880   Min.   :5.070   Min.   :5.100
 1st Qu.:5.300   1st Qu.:5.340   1st Qu.:5.340
 Median :5.460   Median :5.470   Median :5.460
 Mean   :5.448   Mean   :5.482   Mean   :5.483
 3rd Qu.:5.610   3rd Qu.:5.620   3rd Qu.:5.625
 Max.   :5.850   Max.   :5.880   Max.   :5.850
                                 NA's   :6
```

4. Suppose we roll a dice five times. Let Y be the sum of the outcomes in each roll.Find the distribution of Y.

Solution: 4

Range(Y) = {5, 6, 7,....., 29, 30}
Let, X_j represents the outcome of j^{th} roll of a fair die.
The probability distribution of X_j is given by:

$$f(x) = \begin{cases} \frac{1}{6} & \text{for x = 1,2,3,4,5,6} \\ 0 & \text{otherwise} \end{cases}$$

Let, $S_j = X_1 + X_2 + ... + X_j$ = sum of j dice, with probability distribution:
$f_j(y) = P(S_j = y)$; S_j= j, j+1,....., 6j

$$f_j(y) = \sum_{x=1}^{6} P(S_{j-1} = y - x, X_j = x)$$

$$f_j(y) = \frac{1}{6} \sum_{x=1}^{6} P(S_{j-1} = y - x)$$

Let Y= S_5 (putting j=5), then the distribution of Y is given by:

$$P(Y = y) = f_5(y) = \sum_{x=1}^{6} P(S_4 = y - x, X_5 = x)$$

$$P(Y=y)=\frac{1}{6}\sum_{x=1}^{6}f_4(y-x) \quad ; \quad y=5,6,....,30$$

(The method discussed above is known as recursion)
As an example,

$$P(Y=5)=\frac{1}{6}\sum_{x=1}^{6}f_4(5-x)$$

$$P(Y=5)=\frac{1}{6}\left(f_4(5-1)+f_4(5-2)+f_4(5-3)+f_4(5-4)+f_4(5-5)+f_4(5-6)\right)$$

$$P(Y=5)=\frac{1}{6}\left(f_4(4)\right)$$

Because,

$$f_{\mathrm{j}}(x)=0 \quad if \quad x<j$$

Similarly,

$$f_4(4)=\frac{1}{6}\left(f_3(3)\right)=\frac{1}{6^4}$$

Hence,

$$P(Y=5)=\frac{1}{6}\left(f_4(4)\right)=\frac{1}{6}\left(\frac{1}{6^4}\right)=\frac{1}{6^5}$$

5. Toss a fair coin: if head roll a 1-6 flat die (i.e 1,6 have probability $\frac{1}{4}$ and 2,3,4,5 have probability $\frac{1}{8}$); and if tail roll a 3-4 flat die (i.e 3,4 have probability $\frac{1}{4}$ and 1,2,5,6 have probability $\frac{1}{8}$). Let X be the outcome of the toss of a coin. Let Y be the outcome of the roll of the die.

   (a) Find the conditional distribution of $Y|X=Head$

   (b) Find the conditional distribution of $Y|X=Tail$

   (c) Find the $P(X=Head|Y=3)$

Solution: 5
X = Outcome of the toss of a coin.
Y = Outcome of the roll of the die.

(a). The conditional distribution of Y|X=Head is:

$$P(Y=y|X=Head)=\begin{cases}\frac{1}{4} & \text{for } y=1,6\\ \frac{1}{8} & \text{for } y=2,3,4,5\end{cases}$$

As, if head appears on the coin, we are rolling a 1-6 flat die.

(b). The conditional distribution of Y|X=Tail is:

$$P(Y=y|X=Tail)=\begin{cases}\frac{1}{4} & \text{for } y=3,4\\ \frac{1}{8} & \text{for } y=1,2,5,6\end{cases}$$

As, if tail appears on the coin, we are rolling a 3-4 flat die.

(c). The probability of X=Head|Y=3 is given by:

$$
\begin{aligned}
P(X = Head|Y = 3) &= \frac{P(X = Head, Y = 3)}{P(Y = 3)} \\
&= \frac{P(Y = 3|X = Head)P(X = Head)}{P(Y = 3)} \\
&= \frac{P(Y = 3|X = Head)P(X = Head)}{P(Y = 3|X = Head)P(X = Head) + P(Y = 3|X = Tail)P(X = Tail)} \\
&= \frac{\frac{1}{8} \times \frac{1}{2}}{\frac{1}{8} \times \frac{1}{2} + \frac{1}{4} \times \frac{1}{2}} \\
\therefore P(X = Head|Y = 3) &= \frac{1}{3}
\end{aligned}
$$