CS544 Module 2 Assignment

Part1) Probability - 25 points

Use the Bayes theorem to calculate the following probabilities. Show the individual steps of the Bayes theorem. You can use R for the calculations.

Suppose that in a particular state, among 10000 people surveyed, 4250 people are in the age group *18-34* years, 2850 people are in the age group *35-49* years, 1640 people are in the age group *50-64* years, and the remaining are *65 years & over*.

Out of those in the age group *18-34* years, 1062 people had a BMI of above 30. Of those in the age group *35-49* years, 1710 people had a BMI of above 30. Among those in the *50-64* years range, 656 people had a BMI of above 30. In the last age group, 189 people had a BMI of above 30.

a) What is the probability that a randomly selected person in this survey will have a BMI of above 30?

b) If a randomly selected person had a BMI of above 30, what is the probability of that person being in the age group 18-34 years?

c) If a randomly selected person had a BMI of above 30, what is the probability of that person being in the age group 35-49 years?

d) If a randomly selected person had a BMI of above 30, what is the probability of that person being in the age group 50-64 years?

e) If a randomly selected person had a BMI of above 30, what is the probability of that person being in the 65 years & over?

Part2) Random Variables - 25 points

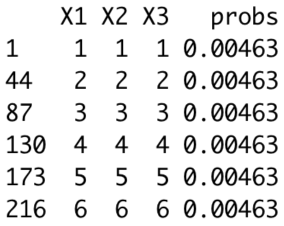
Consider a game which involves rolling **three** dice. Write the **R code** for the following.

Using the **rollDie** function from the **prob** library, setup the sample space for this experiment with the probability space.
For each of the following scenarios from **a)** through **e)**, show the corresponding **outcomes** and the **probability** of that event. The sample outputs for **b)** are shown as example.

a) The sum of the rolls is greater than 10.

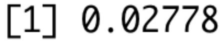
b) All the three rolls are identical.

Sample Output for outcomes:



```
    X1 X2 X3   probs
1    1  1  1 0.00463
44   2  2  2 0.00463
87   3  3  3 0.00463
130  4  4  4 0.00463
173  5  5  5 0.00463
216  6  6  6 0.00463
```

Sample Output for probability:



```
[1] 0.02778
```

c) Only two of the three rolls are identical.

d) None of the three rolls are identical.

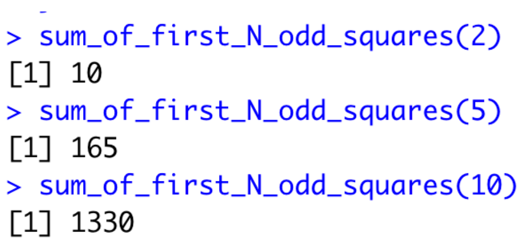
e) Only two of the three rolls are identical given that the sum of the rolls is greater than 10.

Part3) Functions - 20 points

Using a **for** loop or a **while** loop, write your own **R function**,
sum_of_first_N_odd_squares (***n***),
that returns the sum of the squares of the first **n** odd numbers.

For example, if n = 5, the first five odd numbers are 1, 3, 5, 7, 9 and the required result is
$1^2 + 3^2 + 5^2 + 7^2 + 9^2 = 165$.

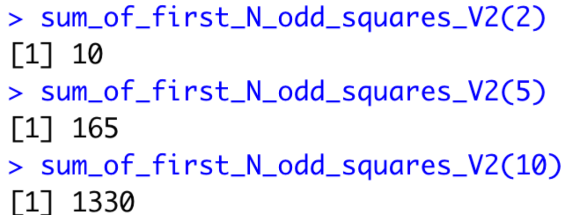
Test your function as follows:



```
> sum_of_first_N_odd_squares(2)
[1] 10
> sum_of_first_N_odd_squares(5)
[1] 165
> sum_of_first_N_odd_squares(10)
[1] 1330
```

Now, **without** using any loop, write your own **R function**,
sum_of_first_N_odd_squares_V2 (***n***),
that returns the sum of the squares of the first **n** odd numbers.

Test your function as follows:



```
> sum_of_first_N_odd_squares_V2(2)
[1] 10
> sum_of_first_N_odd_squares_V2(5)
[1] 165
> sum_of_first_N_odd_squares_V2(10)
[1] 1330
```

Part4) R - 30 points

Initialize the Dow Jones Industrials daily closing data, *dow*, using the read.csv function with the link: http://people.bu.edu/kalathur/datasets/DJI_2020.csv

The first 6 rows of the dataset are as shown below:

```
> head(dow)
    Date Close
1 1/2/20 28869
2 1/3/20 28635
3 1/6/20 28703
4 1/7/20 28584
5 1/8/20 28745
6 1/9/20 28957
```

Provide the simplest R code and output for all of the following. **The code should work for any given data**.

a) Store the result of the **summary** function for the *Close* attribute as the variable *sm*. Change the *names* of this variable so that the output appears as shown below.

```
> sm
  Min    Q1    Q2  Mean    Q3   Max 
18592 23466 24826 25544 28862 29551 
```

Using the above data, show the quartile variations for the 4 quartiles as shown below. You can use **paste** or **sprintf.**

```
[1] "First Quartile variation is 4873.5"
[2] "Second Quartile variation is 1360.5"
[3] "Third Quartile variation is 4035.5"
[4] "Fourth Quartile variation is 689.5"
```

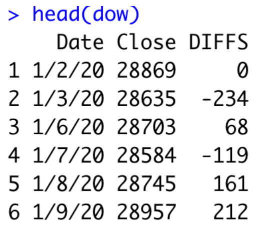
b) Produce the output for the minimum of the Dow closing value in the dataset as shown below:

```
[1] "The minimum Dow value of 18592 is at row 56 on 3/23/20"
```

c) Suppose you have an index fund tied to the Dow closing value. If you have invested on the minimum date, what date from the dataset you would have sold to gain the maximum percentage gain. The output is as shown below. Note that the code should be generic so that it works on any such dataset.

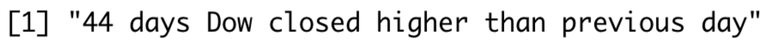
```
[1] "I would sell on 4/29/20 when Dow is at 24634 for a gain of 32.50%"
```

d) Use the **diff** function to calculate the differences between consecutive closing values in the dataset. Insert the value 0 at the beginning of these differences. Add this result as the DIFFS column of the data frame. The result is as shown below.

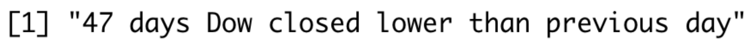


```
> head(dow)
    Date Close DIFFS
1 1/2/20 28869     0
2 1/3/20 28635  -234
3 1/6/20 28703    68
4 1/7/20 28584  -119
5 1/8/20 28745   161
6 1/9/20 28957   212
```

e) How many days did the Dow close higher than its previous day value? How many days did the Dow close lower than its previous day value?

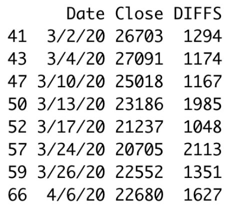


```
[1] "44 days Dow closed higher than previous day"
```



```
[1] "47 days Dow closed lower than previous day"
```

f) Show the subset of the data where there was a **gain** of at least 1000 points from its previous day value.



```
      Date Close DIFFS
41  3/2/20 26703  1294
43  3/4/20 27091  1174
47 3/10/20 25018  1167
50 3/13/20 23186  1985
52 3/17/20 21237  1048
57 3/24/20 20705  2113
59 3/26/20 22552  1351
66  4/6/20 22680  1627
```

Submission:

- You must work on your assignments individually. You are not allowed to copy the answers from the others.
- Each assignment has a strict deadline. However, you are still allowed to submit your assignment within 2 days after the deadline with a penalty. 15% of the credit will be deducted unless you made previous arrangements with your facilitator and professor. Assignments submitted 2 days after the deadline will not be graded.

When the term *lastName* is referenced, please replace it with your last name.

Create a folder, **CS544_HW2_lastName** and place the following files in this folder.

Provide all R code in a single file, **CS544_HW2_lastName.R**. Clearly mark each subpart of each question.

Provide the corresponding code and outputs from the R console in a single Word document, CS544_HW2_lastName.doc.

Archive the folder (**CS544_HW2_lastName.zip**). Upload the zip file to the Assignments section of Blackboard.