CS381/780 Data Analytic Mid-term Exam 3/17/2021

Instruction: For multiple choice questions, clearly circle one of the choice; for all other questions, write your answer right below the questions. All questions carry the same weights.

Name:

Question 1: Given the following database tables, write a SQL statement to list the name of employees in the Finance department who was hired before year 2020

Sample table: employees

emp_id	emp_name	job_name	manager_id	hire_date	salary	commission	dep_id
68319	KAYLING	PRESIDENT		1991-11-18	6000.00		1001
66928	BLAZE	MANAGER	68319	1991-05-01	2750.00		3001
67832	CLARE	MANAGER	68319	1991-06-09	2550.00		1001
65646	JONAS	MANAGER	68319	1991-04-02	2957.00		2001
67858	SCARLET	ANALYST	65646	1997-04-19	3100.00		2001
69062	FRANK	ANALYST	65646	1991-12-03	3100.00		2001
63679	SANDRINE	CLERK	69062	1990-12-18	900.00		2001
64989	ADELYN	SALESMAN	66928	1991-02-20	1700.00	400.00	3001
65271	WADE	SALESMAN	66928	1991-02-22	1350.00	600.00	3001

Sample table: salary_grade

grade	min_sal	max_sal
1	800	1300
2	1301	1500
3	1501	2100
4	2101	3100

Sample table: department

dep_id	dep_name	dep_location
1001	FINANCE	SYDNEY
2001	AUDIT	MELBOURNE
3001	MARKETING	PERTH
4001	PRODUCTION	BRISBANE

Answer:

SELECT emp_name FROM employees e inner join department d on e.dep_id = d.dep_id WHERE d.dep_name = 'FINANCE' and e.hire_date <= '2020-01-01'

Question 2: Based on the database tables in previous question, write a SQL statement to list all employees of grade 3 and 4 whose department location is in Sydney

Answer:

SELECT * FROM employees e, salary_grade s WHERE e.salary BETWEEN s.min_sal AND s.max_sal AND s.grade

IN (3, 4) AND e.emp_id IN (SELECT e.emp_id FROM employees e WHERE e.dep_id in (select dep_id from department where dep_location = 'Sydney'))

Question 3: What does EDA stands for and what are some of the typical tasks in EDA?

Answer:

EDA stands for Exploratory Data Analysis

Question 4: What is the purpose of building a Data warehouse? List out 3 differences between a Data Warehouse and a traditional database?

Answer: DW is an organized collection of integrated, subject-oriented databases designed to support business decision functions. Any of the 3 items in Slide 22 of the DataScience Overview lecure

Question 5: On comparing a typical database in a OLTP system versus a data warehouse in a OLAP system, which of the following are true?

```
1. Data warehouse has more granularity on the transaciton when compared with a
traditional database because it needs more details for business decision support.
2. Updates on a OLTP system typically happens live while updates on OLAP system
can be on a daily, weekly or monthly basis.
3. Data Mart of an enterprise refers to the collection of different Data Warehouse
from different department combined together just like Wal-Mart inventory is coming
from their warehouses located on different regions of the country
```

A. Only 1
B. Only 2
C. 1 and 2
D. 2 and 3
E. 1, 2 and 3

Answer: B (1 is wrong because DW has less granularity, 2 is right, 3 is wrong because DM is on a department level while DW is for the whole enterprise)

Question 6: Who will be responsible for running the ETL process and making sure the results make sense?

A. Data Engineer
B. Data Analyst
C. Data Scientist

Answer: A

Question 7: On given the following weights (lbs) dataset: {80, 100, 100, 80, 110, 70, 90}. Answer the following questions (show your calculation)

A. What is the mean? Answer: 90
B. What is median? Anser: 90
C. What is the mode? Answer: 80 or 100
D. What is the standard deviation? Answer: 13.09

Question 8: What is selection bias and how can you avoid it?

Answer:

Selection bias is the situation where the sample data being not representative of the target population because of the way how the data are selected. Random sampling or stratified sampling are some of the ways to reduce selection bias.

Question 9: Which of the following statements is/are true about Type-1 and Type-2 errors?

```
1. Type-1 error occurs when the prediction is different from the actual case
2. Type-1 error occurs when we fail to reject the null hypothesis while the actual
case is negative
3. Type-2 error occurs when we reject the null hypothesis while the actual case is
positive
```

A. Only 1
B. Only 2
C. 1 and 2
D. 2 and 3
E. None of the above

Answer: E (1 is wrong , 2 is wrong because we had described is a TN case, 3 is true because type-1 is TP)

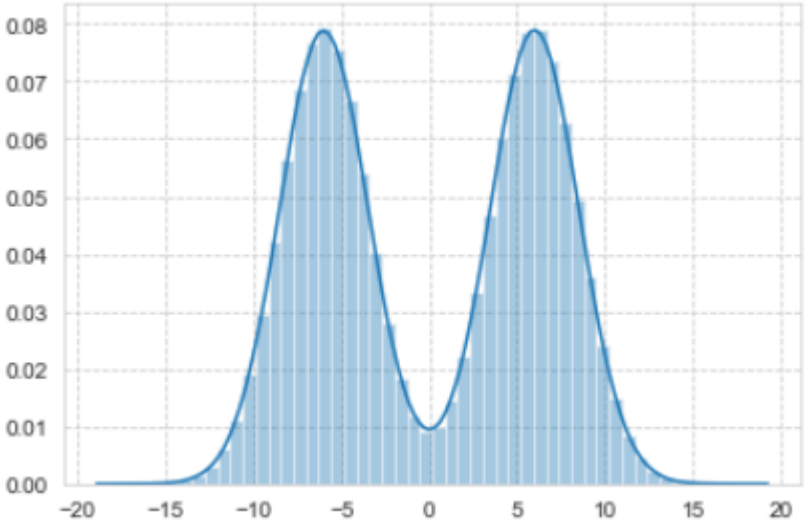
Question 10: Which of the following are true

```
1. In a negatively skewed distribution, the mean will be less than the median
2. In a positively skewed distribution, the median is larger than the mean
3. In a normal distribution, the mean and the mode are the same
```

A. Only 1
B. Only 2
C. Only 3
D. 1 and 3
E. 2 and 3

Answer: D (1 is true, 2 is wrong, 3 is true) (Ref: slide 21 of the Probability and Statistics Review Part I lecture)

Question 11: In the following double hump distribution, which of the following are true




```
1. The skew is zero
2. The mean and the mode are the same
3. The median and mean are the same
```

A. 1 and 2
B. 1 and 3
C. 2 and 3
D. All of the above

Answer: B (1 is true because the distribution is symmetric, 2 is wrong because the mode is the hump, mean is the center, 3 is true)

(Ref: slide 21 of the Probability and Statistics Review Part I lecture)

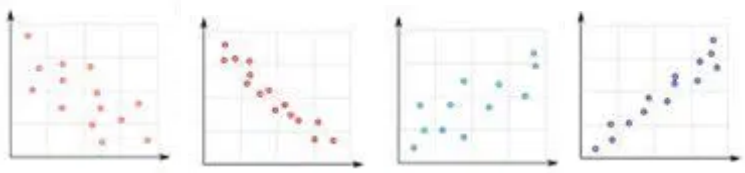
Question 12: Select which of the following are true

```
1. The net worth distribution of a retired community in California will have a
higher mean than the distribution for the whole country
2. The net worth distribution of a retired community in California will have a
higher standard deviation than the distribution for the whole country
3. The net worth distribution of a retired community in California will have a
higher kurtosis than the distribution for the whole country
```

A. Only 1
B. Only 2
C. 1 and 2
D. 1 and 3
E. None of the above

Answer: A (the data are more homogeneous)

Question 13: Suppose you are given the following plots 1-4 (from left to right) and you want to compare their Pearson correlation coefficients. Which of the following are true (including the sign)?



```
1. 1 > 2
2. 3 > 4
3. 2 > 3
```

A. Only 1
B. Only 2
C. 1 and 2
D. 2 and 3
E. None of the above

Answer: A. Say the correlation are -0.8, -1, +0.6, +0.8 then 1 is true, 2 is wrong, 3 is wrong)

Question 14: Name and describe 3 common sampling methods and their corresponding pros and cons.

Answer:

Answer: Random, Systematic, Convenience or Stratified (Ref: slide 10 of the Probability and Statistics Review Part I lecture)

Question 15: In regards to filling missing values, which of the following are true?

```
1. Removing the problematic rows may introduce systematic bias.
2. Using mean to fill in the missing values ia always preferable because it
captures on average the most common possibility.
3. Using median is always preferable than the mean when there are many outliers in
the data set because it is less sensitive.
```

A. Only 1
B. 1 and 2
C. 1 and 3
D. 2 and 3

E. All are correct

Answer: A (1 is true while 2 and 3 are wrong because of the words "always")

Question 16: Explain what is Bias and Variable Trade-off.

Answer: When using a too simplistic model, we will have high bias and low variance while using a too complicated model, our prediction will have low bias and high variance. The first case is underfiting the data while the second is overfitting the data. The consideration in finding the correct balance between bias and variance is referred to as the Bias and Variance trade-off. (Ref: slide 15 to 18 of the Linear Regression lecture)

Question 17: Adding additional variables to a linear regression model will result in

```
1. Increase in R-square
2. Decrease in R-square
3. Increase or Decrease in R-square, depending on the dataset
```

A. Only 1 is correct
B. Only 2 is correct
C. Only 3 is correct
D. None of the above

Answer: A (adding any variable will always increase R-square, slide 14 of the Linear Regression lecture)

Question 18: Suppose you are given a dataset, which of the followings are considered as good common practice in building model.

```
1. Use a big portion of the data points from your dataset to fit your model, and
reserve the rest for testing your model.
2. Take different samples from your dataset to build different versions of a model
may not be a good idea because you will not be sure which one of these models will
be right.
3. Use as many features as possible from your dataset because different features
may be able to explain the different part of the behavior of the target variables.
```

A. Only 1
B. Only 2
C. 1 and 2
D. 1 and 3
E. 1, 2 and 3

Answer: A (1 is true, 2 is wrong as it is cross validation 3 is wrong because using too many features can overfit and some of them may be collinear)

Question 19: In hypothesis testing, which of the following are true

```
1. The null hypothesis will be rejected if the p-value is 0.03 in a two-tail tests
at 95% significance.
2. The higher the p-value, the higher the chance that the null hypothesis will be
rejected.
3. The alternative hyopthesis will be accepted if the probability of the
observation is extremely small.
```

A. Only 1
B. Only 2
C. Only 3
D. 1 and 2
E. 2 and 3

Answer: C (1 is wrong because we need 2.5% on each side of a two-tail test, 2 is wrong because null should be accepted and 3 are true)

Question 20: In regards to Linear and Logistic Regression, which of the followings are true?

```
   1. When we are presented 10 different models with different number and choices
of independent variables in Linear Regression, we always want to pick the model
that has highest R-square as R-square determines how good the fit is.
   2. The range of an odd of an event is between 0 and 1 because by definition
probability has to be between 0 and 1.
   3. R-square is used as a model performance metrics for both Linear and Logistic
Regression.
```

A. Only 1
B. Only 2
C. 1 and 2
D. 1 and 3
E. None of the above

Answer: E (3 is wrong because R-square is used for Linear Regression only, 1 is wrong because of the possibility of overfitting, 2 is wrong because odd is the ratio of probability)