CS381 Data Analytic Mid-term Exam 10/23/2019

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 select customers who worked with a salesman located in Paris

Sample table: salesman

salesman_id	name	city	commission
5001	James Hoog	New York	0.15
5002	Nail Knite	Paris	0.13
5005	Pit Alex	London	0.11
5006	Mc Lyon	Paris	0.14
5007	Paul Adam	Rome	0.13
5003	Lauson Hen	San Jose	0.12

Sample table: customer

customer_id	cust_name	city	grade	salesman_id
3002	Nick Rimando	New York	100	5001
3007	Brad Davis	New York	200	5001
3005	Graham Zusi	California	200	5002
3008	Julian Green	London	300	5002
3004	Fabian Johnson	Paris	300	5006
3009	Geoff Cameron	Berlin	100	5003
3003	Jozy Altidor	Moscow	200	5007

Answer: select customer.cust_name as "customer", salesman.name as "salesman", salesman.city from salesman, customer where salesman.salesman_id = customer.salesman_id and salesman.city = 'Paris'

Question 2: Given the following database tables, Write a SQL statement to select employees in the IT department who earn more than 30000

Sample table: departments

DEPARTMENT_ID	DEPARTMENT_NAME	MANAGER_ID	LOCATION_ID
10	Administration	200	1700
20	Marketing	201	1800
30	Purchasing	114	1700
40	Human Resources	203	2400
50	Shipping	121	1500
60	IT	103	1400
70	Public Relations	204	2700
80	Sales	145	2500

Sample table: employees

EMPLOYEE_ID	FIRST_NAME	LAST_NAME	EMAIL	PHONE_NUMBER	HIRE_DATE	JOB_ID	SALARY	COMMISSION_PCT	MANAGER_ID	DEPARTMENT_ID
100	Steven	King	SKING	515.123.4567	2003-06-17	AD_PRES	24000.00	0.00	0	90
101	Neena	Kochhar	NKOCHHAR	515.123.4568	2005-09-21	AD_VP	17000.00	0.00	100	90
102	Lex	De Haan	LDEHAAN	515.123.4569	2001-01-13	AD_VP	17000.00	0.00	100	90
103	Alexander	Hunold	AHUNOLD	590.423.4567	2006-01-03	IT_PROG	9000.00	0.00	102	60
104	Bruce	Ernst	BERNST	590.423.4568	2007-05-21	IT_PROG	6000.00	0.00	103	60
105	David	Austin	DAUSTIN	590.423.4569	2005-06-25	IT_PROG	4800.00	0.00	103	60
106	Valli	Pataballa	VPATABAL	590.423.4560	2006-02-05	IT_PROG	4800.00	0.00	103	60

Answer: select e.FIRST_NAME, e.LAST_NAME, d.DEPARTMENT_NAME, e.SALARY from departments as d, employees as e where e.DEPARTMENT_ID = d.DEPARTMENT_ID and e.SALARY > 30000

Question 3: What does OLAP and OLTP stands for?

Answer: On-line Transaction Processing and On-Line Analytical Processing (Ref: slide 17 of the Data Science Overview lecture)

Question 4: What is the difference between a Data Mart versus a Data Warehouse?

Answer: Data Warehouse is for the whole enterprise while Data Mart is a smaller version of a Data Warehouse and is tailored towards a more specific goal or for a smaller department. (Ref: slide 19 of the Data Science Overview lecture)

Question 5: Which of the following are true?

```
1. Data Warehouse is built for storing operation transactions efficiently
2. Data Warehouse contains lower granularity data when compared with database
3. Star schema is used more frequently used in a Data Warehouse than in database
```

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

Answer: D, 1 is for Database (Ref: slide 22 of the Data Science Overview lecture)

Question 6: Who will be responsible for collecting unstructured data and transform them into database table?

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

Answer: A (Ref: slide 15 of the Data Science Overview lecture)

Question 7: Given the following weights (lbs) dataset: { 120, 140, 140, 120, 150, 110, 130}.

A. What is the mean: Your answer: 130
B. What is the median: Your answer: 130
C. What is the mode: Your answer: 120 or 140

Ref: slide 15 of the Probability and Statistic Review, Part I lecture)

Question 8: What is Central Limit Theorem?

Answer: The Central Limit Theorem states that regardless of the shape of the population distribution, the distribution of sample means will be approximately normal (Ref: slide 23 of the Probability and Statistic Review, Part I lecture) http://www.statisticslectures.com/topics/centrallimittheorem/

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

```
1. Type-1 error occurs when we rejects a null hypothesis when it is actually true
2. Type-2 error occurs when the prediction is positive while the actual case is
negative
3. Type-2 error means the prediction is wrong from the actual case
```

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

Answer: 1 is saying Type 1 error is FP which is true, 2 is saying Type-2 error is FP (therefore it is wrong), 3 is just wrong so, the answer is A

Question 10: Which of the following are true

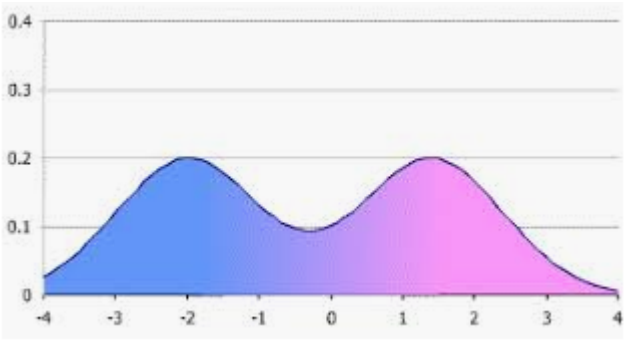
```
1. In a positively skewed distribution, the median is larger than the mean
2. In a negatively skewed distribution, the mean will be less than the median
3. In a normal distribution, the mean and the mode are the same
```

A. Only 1
B. Only 2

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

Answer: E (1 is wrong, 2 is true, 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 median and mean are the same
2. The skew is zero
3. The mean and the mode are the same
```

A. Only 1
B. 1 and 3
C. 1 and 2
D. All are true

Answer: C (2 is true because the distribution is symmetric, 3 is wrong because the mode is the hump, mean is the center) (Ref: slide 21 of the Probability and Statistics Review Part I lecture)

Question 12: Select which of the following are true

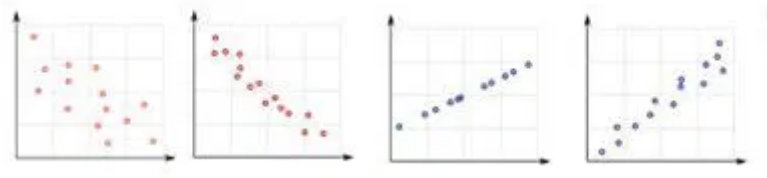
```
1. Income distribution of a gated community will have a higher mean than the
distribution for the whole country
2. Income distribution of a gated community will have a higher standard deviation
than the distribution for the whole country
3. Income distribution of a gated community will have a higher 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 (Think in term of whehter the distrbution is more heterogeneous or more homogeneous, in a more homogeneous dataset, in the extreme case, every data point is the same, then the standard deviation is zero

and there is almost no outliers and therefore it has low standard deviation and low kurtosis)

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. 3 > 2
```

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

Answer: Say the correlation are -0.8, -1, 1, +0.8, then 1 is wrong, 2 is true, 3 is true. So the answer is D

Question 14: Name 3 common sampling methods

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

Question 15: What does EDA stands for and what are some of the common tasks in EDA?

Answer: Exploratory Data Analysis. Filling missing values, removing outliers, calculating basic statistics, plot some graphs (Ref: slide 1 of the Exploratory Data Analysis lecture)

Question 16: When filling missing values, what is the reason for using median instead of the mean? Under what other situations that you prefer using the mean instead of median?

Answer: Use median when we do not want to be too sensitive to outliers. Use mean when we do want to include effects from most outliers . (Ref: slide 4 of the Exploratory Data Analysis lecture)

Question 17: Explain what is the Bias and Variance 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 seconde is overfitting the data. To find the correct balance is referred to as the Bias and Variance trade-off. (Ref: slide 15 to 18 of the Linear Regression lecture)

Question 18: Adding a non-important feature 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. Either 1 or 2 is correct
E. None of the above

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

Question 19: What is the purpose of running a K-fold cross validation?

Answer: To test how robust the model is with respect to different samples of training dataset

Question 20: Answer the following questions based on the following results from a classification model

n= 100	**Predicted: NO**	**Predicted: YES**
Actual: NO	25	7
Actual: YES	8	60

A. Calculate the Precision: Your answer: TP/(TP+FP) = 60/67
B. Calculate the Recall: Your answer: TP/(TP+FN) = 60/68
C. Calculate the True Positive Rate: Your answer: Recall = TP/(TP+FN) = 60/68
D. Calculate the False Positive Rate: Your answer: FP/(FP+TN) = 7/32