1- What is the relationship between similarity and distance, how do you formulate it (3 points)

**similarity = 1 – distance**

**they are opposite**

2- Which of the following correlations are more useful? (3 points)

**a-b ==> -0.95**

a-c ==> 0.1

**a-d ==> 0.70**

a-e ==> -0.07

3- Write the formula of scaling and standardizing (4 points)

**Scaling (0-1)**

**df[‘age’] = (df[‘age’] – df[‘age’].min()) / (df[‘age’].max() - df[‘age’].min())**

**Standardizing**

**df[‘age’] = (df[‘age’] – df[‘age’].mean()) / df[‘age’].std()**

4- Compare the jaccard distance and euclidean distance, where they can be used, what are their problems? (12 points)

**Jaccard: multiple categorical variables (a: [milk, egg, bread], b: [milk, battery, bread] )**

**weakness: TF-IDF ==> frequency**

**Euclidean: distance of two points: kusbakisi**

**weakness: need to normalize, need to scale according to dimension**

**age, salary**

5- Explain 2 filling methods and compare them, explain if there is a weakness (6 points)

**fill with “zero”: mean changes! [age, empty]**

**fill with “median”: standard deviation gets smaller!!!, data shape changes**

**fill with “average”: standard deviation gets smaller!!!, data shape changes**

**fill with “previous”: prev customer do not need to be same as the next customer**

6- You are working in a company and your duty is to find the main reasons of customer complainments, what do you do? (13 points)

**Represent the sentences as words!. Find the most used words!**

**Jaccard distance can be used**

7- Give an example of a business problem which can be solved with “two” different problem domains (classification, ….) (10 points)

**churn: [classification 1/0, regression: how much they churned, clustering: 5 group ]**

**Exercise more**

8- What is heatmap, what is it used for (3 points)

**Visualize two categorical variables. Heat = sicaklik = renk**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Medium** | **Small** | **...c** | **d** |
| **Baku** | **15** | **60** |  |  |
| **Istanbul** |  |  |  |  |
| **r** |  |  |  |  |
| **s** |  |  |  |  |

**Crosstab**

9- Write a simple function to find mean of a list without using pandas or numpy (3 points)

**lst = [1, 2, 3, 4]**

**sum( lst ) / len( lst )**

10- Write a simple function to generate a random number with given range (3 points)

**import random**

**random.randint( 100, 300 )**

**100+random.random() \* (300-100)**

**0-1**

**100-300**

11- Write the types of data (categoric, flag, numeric etc) below (3 points)

**gender : flag**

**age : numeric**

**birth month : categoric**

**salary : numeric**

**education level : categoric(!) ==> ordinal (high school < university, university < master, master < doctorate)**

12- Analyze the dataset and extract some useful insights (25 points)

13- Which features are more important to determine “Response” (12 points)

correlation

**Previously\_Insured**

**Vehicle\_Age**

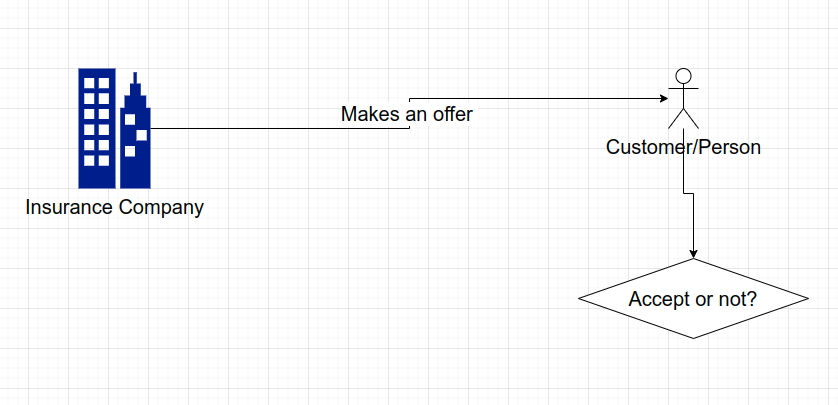
**Vehicle\_Damage**

=====

Our company is a vehicle Insurance company that has provided Insurance to its customers.

They do want to use machine learning to predict next years sales whether which customers will accept the offer or not .

So the company will make an offer of “Anuual Premium” to its customers, the task is to analyze dataset for predicting Response from the customer.



Dataset

| Variable | Definition |
| --- | --- |
| id | Unique ID for the customer |
| Gender | Gender of the customer |
| Age | Age of the customer |
| Driving\_License | 0 : Customer does not have DL, 1 : Customer already has DL |
| Region\_Code | Unique code for the region of the customer |
| Previously\_Insured | 1 : Customer already has Vehicle Insurance, 0 : Customer doesn't have Vehicle Insurance |
| Vehicle\_Age | Age of the Vehicle |
| Vehicle\_Damage | 1 : Customer got his/her vehicle damaged in the past. 0 : Customer didn't get his/her vehicle damaged in the past. |
| Annual\_Premium | The amount customer needs to pay as premium in the year |
| PolicySalesChannel | Anonymized Code for the channel of outreaching to the customer ie. Different Agents, Over Mail, Over Phone, In Person, etc. |
| Vintage | Number of Days, Customer has been associated with the company |
| Response | 1 : Customer is interested, 0 : Customer is not interested |

Assignment for this weekend (week6):

From the given dataset (in quiz), cleanup, remove outliers, filter the dataset. Make all preparations.

Find the averages of two datasets (divide the dataset into Response=1 and Response=0)

Plot the two datasets and find their similar and different columns by examining the means, standard deviations, distributions (standard deviation, histogram, skewness ..) [same plot]

Find a similarity metric for the customers (do not use the column Response).

dataset0 = df[ df[‘Response’] == 0 ]

dataset1 = df[ df[‘Response’] == 1 ]

dataset0mean = …..

customerA = [ a customer whose response is 1]

similary( customerA, dataset0mean )

Find the customers who is response = 1, but they are similar to the dataset (response = 0)

Find the customers who is response = 0, but they are similar to the dataset (response = 1)