# Advances in Data Science Final Exam Assignment

**Remarks:**

Write the code yourself. ***Cheating is strictly forbidden.***

For each problem write your code in the function format and give the names of the functions as problem numbers, for example for the solution of problem1:

def problem1(input): return something

Put the codes for all problems into one python file and name that file in the following format: NameSurname\_AdvDataScience\_FinalExam.py

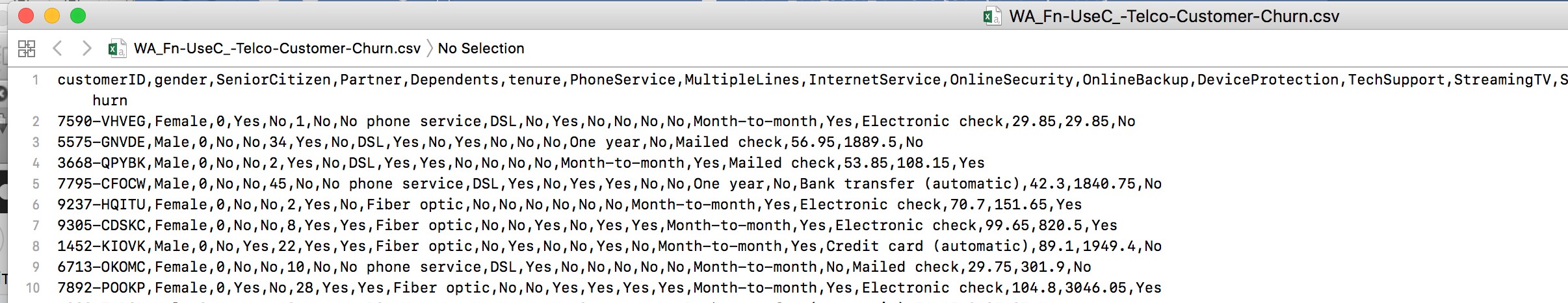
Give as much as documentation for your script using comments.

Please upload your solutions to ninova system. If you are having problems with the ninova system you can send your solutions to [badays@itu.edu.tr](mailto:badays@itu.edu.tr) e-mail address.

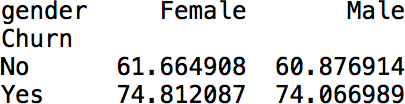
Note: You can use PANDAS or NumPy libraries for the following problems.

# Problem 1 (25 Points).

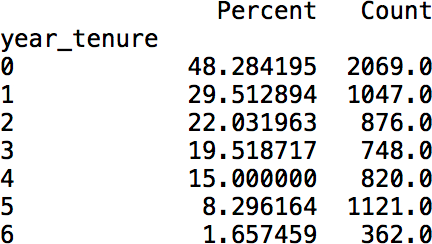
You are given “WA\_Fn-UseC\_-Telco-Customer-Churn.csv” file on Telcom customer churn taken from https://[www.kaggle.com/blastchar/telco-customer-churn.](http://www.kaggle.com/blastchar/telco-customer-churn) You can find more information about the columns in the link provided.



Produce the following table showing the average of Monthly charges grouped by customers who churned or not. The table also is grouped by gender.



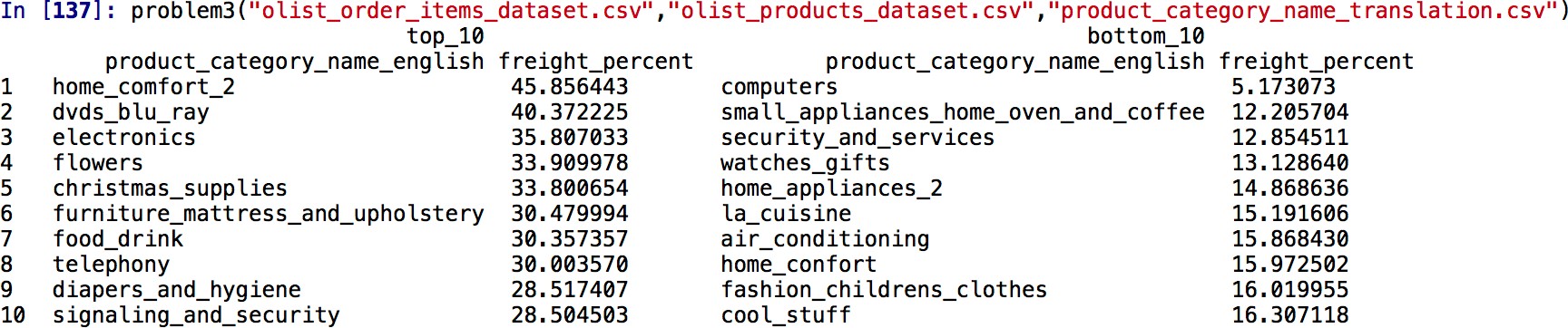
# Problem 2 (25 Points).

For this problem, use the data file given problem1. In that data file, the tenure column shows the number of months the customer has stayed with the company. Here we want to analyse the churn rate as a function of tenure year. Produce the following table for this analysis. Year tenure is zero if the tenure month is less than 12, one if thenure month is between 12 and 23, and so on. In the table, the count column shows the number of data points. The Percent column shows the percentage of the “Yes” in the churn column.

# Problem 3 (25 Points).

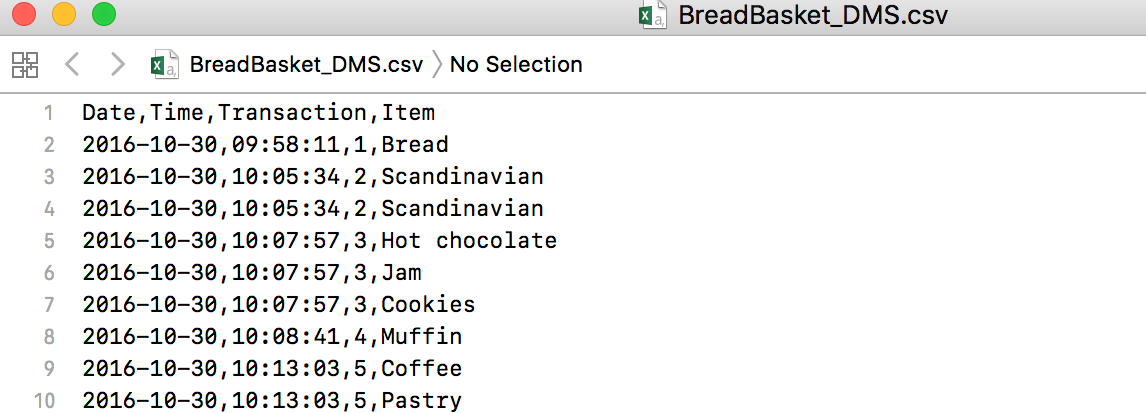
In kaggle.com there is a Brazilian ecommerce public dataset of orders of Olist Store. Go to following link and download the dataset. ht[tps://www.kaggle.com/olistbr/brazilian-ecommerce.](http://www.kaggle.com/olistbr/brazilian-ecommerce) Note that the data is distributed to mulitple csv files. In this problem you need only product\_category\_name\_english.csv, olist\_products\_dataset.csv and olist\_order\_items\_dataset.csv datafiles. Merge these files into one dataframe and do the following analysis.

First calculate the percentage of the freight (delivery) cost to the total cost. The total cost is the sum of freight value and the price. Then calculate the average freight\_percent for each product category. At the final step, report the top-10 and bottom-10 product categories having the heighest and lowest freight percentage, respectively. The final result should be in one table like the following:



# Problem 4 (25 Points).

You are given “BreadBasket\_DMS.csv” file that contains sales informations of a bakery. The columns are: Date,Time,Transaction and Item. Transaction can be thought as the order id.



Using this data, print a table showing the monthly sales quantity of coffee and Tea.

