{"metadata":{"kernelspec":{"language":"python","display_name":"Python 3","name":"python3"},"language_info":{"name":"python","version":"3.7.12","mimetype":"text/x-python","codemirror_mode":{"name":"ipython","version":3},"pygments_lexer":"ipython3","nbconvert_exporter":"python","file_extension":".py"}},"nbformat_minor":4,"nbformat":4,"cells":[{"cell_type":"markdown","source":"# Importing the Libraries","metadata":{}},{"cell_type":"code","source":"!pip install chart_studio","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:54:55.421558Z","iopub.execute_input":"2022-11-20T06:54:55.422273Z","iopub.status.idle":"2022-11-20T06:55:02.230360Z","shell.execute_reply.started":"2022-11-20T06:54:55.422162Z","shell.execute_reply":"2022-11-20T06:55:02.228625Z"},"trusted":true},"execution_count":1,"outputs":[{"name":"stdout","text":"^C\n","output_type":"stream"}]},{"cell_type":"code","source":"# data operation libraries\nimport numpy as np\nimport pandas as pd\n\n# importing visualisation libraries\nimport matplotlib.pyplot as plt\nimport seaborn as sns\n%matplotlib inline\n\n# for chloroplath plotting\nimport chart_studio.plotly as py\nimport plotly.graph_objs as go \nimport plotly\nimport cufflinks as cf\nfrom plotly.offline import download_plotlyjs, init_notebook_mode, plot, iplot\ninit_notebook_mode(connected=True)\ncf.go_offline()\n\n# for datetime operations\nimport datetime as dt\n\n# pandas general settings\npd.options.display.max_columns = None","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:02.233564Z","iopub.execute_input":"2022-11-20T06:55:02.234662Z","iopub.status.idle":"2022-11-20T06:55:03.626296Z","shell.execute_reply.started":"2022-11-20T06:55:02.234596Z","shell.execute_reply":"2022-11-20T06:55:03.619342Z"},"trusted":true},"execution_count":2,"outputs":[{"traceback":["\u001b[0;31m---\u001b[0m","\u001b[0;31mModuleNotFoundError\u001b[0m   Traceback (most recent call last)","\u001b[0;32m/tmp/ipykernel_17/2422436338.py\u001b[0m in \u001b[0;36m<module>\u001b[0;34m\u001b[0m\n\u001b[1;32m   9\u001b[0m \u001b[0;34m\u001b[0m\u001b[0m\n\u001b[1;32m   10\u001b[0m \u001b[0;31m# for chloroplath plotting\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\n\u001b[0;32m---> 11\u001b[0;31m \u001b[0;32mimport\u001b[0m \u001b[0mchart_studio\u001b[0m\u001b[0;34m.\u001b[0m\u001b[0mplotly\u001b[0m \u001b[0;32mas\u001b[0m \u001b[0mpy\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\n\u001b[1;32m   12\u001b[0m \u001b[0;32mimport\u001b[0m \u001b[0mplotly\u001b[0m\u001b[0;34m.\u001b[0m\u001b[0mgraph_objs\u001b[0m \u001b[0;32mas\u001b[0m \u001b[0mgo\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\n\u001b[1;32m   13\u001b[0m \u001b[0;32mimport\u001b[0m \u001b[0mplotly\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\u001b[0;34m\u001b[0m\n","\u001b[0;31mModuleNotFoundError\u001b[0m: No module named 'chart_studio'"],"ename":"ModuleNotFoundError","evalue":"No module named 'chart_studio'","output_type":"error"}]},{"cell_type":"markdown","source":"# Importing the Dataset","metadata":{}},{"cell_type":"code","source":"data = pd.read_csv('../input/global-super-store-dataset/Global_Superstore2.csv', encoding='windows-1252')","metadata":{"scrolled":true,"execution":{"iopub.status.busy":"2022-11-20T06:55:03.627577Z","iopub.status.idle":"2022-11-20T06:55:03.628985Z","shell.execute_reply.started":"2022-11-20T06:55:03.628650Z","shell.execute_reply":"2022-11-20T06:55:03.628685Z"},"trusted":true},"execution_count":null,"outputs":[]},{"cell_type":"markdown","source":"# Data Preparation","metadata":{}},{"cell_type":"code","source":"data.head(2) #taking a look at the dataframe structure","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.630874Z","iopub.status.idle":"2022-11-20T06:55:03.631872Z","shell.execute_reply.started":"2022-11-20T06:55:03.631554Z","shell.execute_reply":"2022-11-20T06:55:03.631585Z"},"trusted":true},"execution_count":null,"outputs":[]},{"cell_type":"code","source":"# correcting 'Order Date' variable\ndata[['order_day','order_month','order_year']] = data['Order Date'].str.split('-', expand=True)\ndata['Order Date'] = data['order_year'] + '/' + data['order_month'] + '/' + data['order_day']\ndata['Order Date'] = pd.to_datetime(data['Order Date'])","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.633741Z","iopub.status.idle":"2022-

11-20T06:55:03.634721Z","shell.execute_reply.started":"2022-11-
20T06:55:03.634395Z","shell.execute_reply":"2022-11-
20T06:55:03.634426Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"# doing likewise for 'Ship
Date'\ndata[['ship_day','ship_month','ship_year']] = data['Ship Date'].str.split('-',
expand=True)\ndata['Ship Date'] = data['ship_year'] + '/' + data['ship_month'] + '/' +
data['ship_day']\ndata['Ship Date'] = pd.to_datetime(data['Ship Date'])","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.636508Z","iopub.status.idle":"2022-
11-20T06:55:03.637482Z","shell.execute_reply.started":"2022-11-
20T06:55:03.637150Z","shell.execute_reply":"2022-11-
20T06:55:03.637182Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"# dropping the support columns\ndata.drop(columns=
['order_day','order_month','order_year','ship_day','ship_month','ship_year'],
inplace=True)","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.639014Z","iopub.status.idle":"2022-11-
20T06:55:03.639727Z","shell.execute_reply.started":"2022-11-
20T06:55:03.639513Z","shell.execute_reply":"2022-11-
20T06:55:03.639535Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"data.info() #checkout the data types/ null rows and memory
consumption","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.664226Z","iopub.status.idle":"2022-11-
20T06:55:03.665277Z","shell.execute_reply.started":"2022-11-
20T06:55:03.664948Z","shell.execute_reply":"2022-11-
20T06:55:03.664980Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"# let's check out the columns which are suitable category
column type\n\ndata.nunique()","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.667162Z","iopub.status.idle":"2022-11-
20T06:55:03.668164Z","shell.execute_reply.started":"2022-11-
20T06:55:03.667851Z","shell.execute_reply":"2022-11-
20T06:55:03.667882Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"data['Ship Mode'] = data['Ship
Mode'].astype('category')\ndata['Segment'] =
data['Segment'].astype('category')\ndata['Country'] =
data['Country'].astype('category')\ndata['Market'] =
data['Market'].astype('category')\ndata['Region'] =
data['Region'].astype('category')\ndata['Category'] =
data['Category'].astype('category')\ndata['Sub-Category'] = data['Sub-
Category'].astype('category')\ndata['Order Priority'] = data['Order
Priority'].astype('category')","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.669992Z","iopub.status.idle":"2022-11-
20T06:55:03.670982Z","shell.execute_reply.started":"2022-11-
20T06:55:03.670672Z","shell.execute_reply":"2022-11-
20T06:55:03.670703Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"data.info() #check the reduction in memory
consumption","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.672803Z","iopub.status.idle":"2022-11-
20T06:55:03.673817Z","shell.execute_reply.started":"2022-11-
20T06:55:03.673509Z","shell.execute_reply":"2022-11-
20T06:55:03.673541Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"# making sure neither of our category columns have leading
spaces\n\ndef remove_leading_spaces(df):\n   for cols in df.columns:\n   if
df[cols].dtypes in ['object','category']:\n   df[cols] = df[cols].str.strip()\n
return df","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.675634Z","iopub.status.idle":"2022-11-
20T06:55:03.676590Z","shell.execute_reply.started":"2022-11-
20T06:55:03.676256Z","shell.execute_reply":"2022-11-
20T06:55:03.676285Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"data = remove_leading_spaces(data)","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.678390Z","iopub.status.idle":"2022-11-
20T06:55:03.679362Z","shell.execute_reply.started":"2022-11-
20T06:55:03.679028Z","shell.execute_reply":"2022-11-
20T06:55:03.679058Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"data.head(2)","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.681139Z","iopub.status.idle":"2022-11-
20T06:55:03.682109Z","shell.execute_reply.started":"2022-11-

20T06:55:03.681798Z","shell.execute_reply":"2022-11-
20T06:55:03.681827Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"# generating years from our 'Order_year' variable because we
are going \n# to need this in future analysis\n\ndata['Order_year'] = data['Order
Date'].dt.year","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.683873Z","iopub.status.idle":"2022-11-
20T06:55:03.684833Z","shell.execute_reply.started":"2022-11-
20T06:55:03.684524Z","shell.execute_reply":"2022-11-
20T06:55:03.684555Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"# also total unique customer count is something we need in our
future analysis\n\nprint('Number of unique customers made purchase in 2011:
{}'.format(data[data['Order_year']==2011]['Customer Name'].nunique()))\nprint('Number of
unique customers made purchase in 2012: {}'.format(data[data['Order_year']==2012]['Customer
Name'].nunique()))\nprint('Number of unique customers made purchase in 2013:
{}'.format(data[data['Order_year']==2013]['Customer Name'].nunique()))\nprint('Number of
unique customers made purchase in 2014: {}'.format(data[data['Order_year']==2014]['Customer
Name'].nunique()))","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.686427Z","iopub.status.idle":"2022-11-
20T06:55:03.687633Z","shell.execute_reply.started":"2022-11-
20T06:55:03.687295Z","shell.execute_reply":"2022-11-
20T06:55:03.687346Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"def total_purchase_in_year(row):\n   Order_year = row[24]\n
\n   if Order_year in [2011,2012,2013]:\n   return 795\n   else:\n   return
794\n   \n   \n# generating 'unique_customers_within_year' based on associated year
value\n# for that particular row\n\ndata['unique_customers_within_year'] =
data.apply(total_purchase_in_year, axis='columns')","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.689449Z","iopub.status.idle":"2022-11-
20T06:55:03.690393Z","shell.execute_reply.started":"2022-11-
20T06:55:03.690065Z","shell.execute_reply":"2022-11-
20T06:55:03.690095Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"Before generating revenue column let's understand the
intution behing Revenue.\n\n>Revenue is another word for the amount of money a company
generates from its sales.\n\n>Revenue is most simply calculated as the number of units sold
multiplied by the selling price.\n\n\n<img src='https://stockanalysis.com/img/term/revenue-
formula.png' width='300'/>","metadata":{}},{"cell_type":"code","source":"# Generating
'Revenue' column\ndata['Revenue'] = data['Sales'] * data['Quantity']","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.692173Z","iopub.status.idle":"2022-
11-20T06:55:03.693119Z","shell.execute_reply.started":"2022-11-
20T06:55:03.692810Z","shell.execute_reply":"2022-11-
20T06:55:03.692840Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# EDA\n\nSolving the questions we have been
asked","metadata":{}},{"cell_type":"markdown","source":"# Customers Analysis\n\n**Question:
1**Profile the customers based on their frequency of purchase - calculate frequency of
purchase for each customer\n\n**Question: 2**Do the high frequent customers are contributing
more revenue\n\n**Question: 3**Are they also profitable - what is the profit margin across
the buckets\n\n**Question: 4**Which customer segment is most profitable in each
year.\n\n**Question: 5**How the customers are distributed across the countries- -
","metadata":{}},{"cell_type":"markdown","source":"# Question:1","metadata":{}},
{"cell_type":"markdown","source":"Profile the customers based on their frequency of purchase
- calculate frequency of purchase for each customer","metadata":{}},
{"cell_type":"code","source":"purchase_frequency = data.groupby(['Order_year','Customer
Name'])","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.694915Z","iopub.status.idle":"2022-11-
20T06:55:03.695884Z","shell.execute_reply.started":"2022-11-
20T06:55:03.695572Z","shell.execute_reply":"2022-11-
20T06:55:03.695603Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"purchase_frequency.agg({'Customer Name': 'count',\n
'unique_customers_within_year': 'min',\n   'Revenue': 'sum',\n
'Profit': 'sum'}) ","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.697690Z","iopub.status.idle":"2022-11-
20T06:55:03.698658Z","shell.execute_reply.started":"2022-11-
20T06:55:03.698335Z","shell.execute_reply":"2022-11-
20T06:55:03.698367Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"analysis_result = purchase_frequency.agg({'Customer Name':
'count',\n   'unique_customers_within_year': 'min',\n

'Revenue': 'sum',\n   'Profit': 'sum'})","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.700477Z","iopub.status.idle":"2022-11-
20T06:55:03.701446Z","shell.execute_reply.started":"2022-11-
20T06:55:03.701114Z","shell.execute_reply":"2022-11-
20T06:55:03.701144Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"analysis_result.rename(mapper={'Customer Name':
'Purchase_during_year'}, axis=1, inplace=True)","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.703188Z","iopub.status.idle":"2022-11-
20T06:55:03.704157Z","shell.execute_reply.started":"2022-11-
20T06:55:03.703852Z","shell.execute_reply":"2022-11-
20T06:55:03.703881Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"Calculating Customer Purchase Frequency\n\n\nThe repeat
purchase rate is a calculation that shows you the percentage of your current customer base
that has purchased at least a second time in a specific duration (usally take 365 days).
This metric is influenced by your customer retention efforts and is a good indicator of the
value you are providing your customers.\n\n<img
src='https://blog.smile.io/content/images/2020/05/How-to-Calculate-Purchase-Frequency---RPR-
equation.png' width='300'/>","metadata":{}},
{"cell_type":"code","source":"analysis_result['Customer_purchase_frequency'] =
analysis_result['Purchase_during_year']/analysis_result['unique_customers_within_year']
*100","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.705888Z","iopub.status.idle":"2022-11-
20T06:55:03.706852Z","shell.execute_reply.started":"2022-11-
20T06:55:03.706551Z","shell.execute_reply":"2022-11-
20T06:55:03.706581Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer:\n\nHere we are only supposed to find the
purchase frequency of each customer and not the one who are having the highest purchase
frequency. So here is the result:","metadata":{}},
{"cell_type":"code","source":"analysis_result.head(5)","metadata":
{"scrolled":true,"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.708768Z","iopub.status.idle":"2022-11-
20T06:55:03.709751Z","shell.execute_reply.started":"2022-11-
20T06:55:03.709429Z","shell.execute_reply":"2022-11-
20T06:55:03.709459Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Question:2\n\nDo the high frequent customers are
contributing more revenue?\n\n\nThe question here is comapring the high purchase frequency
customers with high revenue generating customers. In the previous question we found out the
purchase frequency of each customer, so out of those we will finf out highest purchse
frequency customers for that year and then will compare to the highest revenue generator for
that year.","metadata":{}},{"cell_type":"code","source":"tmp_df =
analysis_result.reset_index()","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.711478Z","iopub.status.idle":"2022-11-
20T06:55:03.712205Z","shell.execute_reply.started":"2022-11-
20T06:55:03.711969Z","shell.execute_reply":"2022-11-
20T06:55:03.711990Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"tmp_df.head()","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.763293Z","iopub.status.idle":"2022-11-
20T06:55:03.764641Z","shell.execute_reply.started":"2022-11-
20T06:55:03.764229Z","shell.execute_reply":"2022-11-
20T06:55:03.764264Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"grouped_object = tmp_df.groupby(['Order_year'])","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.767141Z","iopub.status.idle":"2022-
11-20T06:55:03.768163Z","shell.execute_reply.started":"2022-11-
20T06:55:03.767839Z","shell.execute_reply":"2022-11-
20T06:55:03.767870Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"freq_df = pd.DataFrame(columns=tmp_df.columns)","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.770254Z","iopub.status.idle":"2022-
11-20T06:55:03.771267Z","shell.execute_reply.started":"2022-11-
20T06:55:03.770949Z","shell.execute_reply":"2022-11-
20T06:55:03.770981Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"for g,d in grouped_object:\n   highest_freq_customers =
d.nlargest(1, 'Customer_purchase_frequency')\n   freq_df = pd.concat([freq_df,
highest_freq_customers])","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.773369Z","iopub.status.idle":"2022-11-
20T06:55:03.774418Z","shell.execute_reply.started":"2022-11-

20T06:55:03.774084Z","shell.execute_reply":"2022-11-
20T06:55:03.774115Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"def highlight_cols(x): \n   df = x.copy()\n   df.loc[:,
['Customer Name','Customer_purchase_frequency']] = 'background-color: green'\n
df[['Order_year','Purchase_during_year','unique_customers_within_year','Revenue','Profit']]
= 'background-color: grey'\n   return df ","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.776234Z","iopub.status.idle":"2022-11-
20T06:55:03.777231Z","shell.execute_reply.started":"2022-11-
20T06:55:03.776919Z","shell.execute_reply":"2022-11-
20T06:55:03.776952Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"display(freq_df.style.apply(highlight_cols, axis =
None))","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.779040Z","iopub.status.idle":"2022-11-
20T06:55:03.780024Z","shell.execute_reply.started":"2022-11-
20T06:55:03.779716Z","shell.execute_reply":"2022-11-
20T06:55:03.779746Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"rev_df = pd.DataFrame(columns=tmp_df.columns)","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.781864Z","iopub.status.idle":"2022-
11-20T06:55:03.782837Z","shell.execute_reply.started":"2022-11-
20T06:55:03.782527Z","shell.execute_reply":"2022-11-
20T06:55:03.782558Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"for g,d in grouped_object:\n   highest_rev_customers =
d.nlargest(1, 'Revenue')\n   rev_df = pd.concat([rev_df,
highest_rev_customers])","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.784689Z","iopub.status.idle":"2022-11-
20T06:55:03.785647Z","shell.execute_reply.started":"2022-11-
20T06:55:03.785326Z","shell.execute_reply":"2022-11-
20T06:55:03.785357Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"def highlight_cols(x): \n   df = x.copy()\n   df.loc[:,
['Customer Name','Revenue']] = 'background-color: green'\n
df[['Order_year','Purchase_during_year','unique_customers_within_year','Profit','Customer_pu
rchase_frequency']] = 'background-color: grey'\n   return df ","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.787482Z","iopub.status.idle":"2022-11-
20T06:55:03.788444Z","shell.execute_reply.started":"2022-11-
20T06:55:03.788115Z","shell.execute_reply":"2022-11-
20T06:55:03.788145Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"display(rev_df.style.apply(highlight_cols, axis =
None))","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.790534Z","iopub.status.idle":"2022-11-
20T06:55:03.791508Z","shell.execute_reply.started":"2022-11-
20T06:55:03.791163Z","shell.execute_reply":"2022-11-
20T06:55:03.791194Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer:\n\n\nWe can clearly see by comparing both the
tables that neither of the high puchase frequency customers are there in the high revenue
generating customer taable. \n\nSo the answer is no, high purchase frequency customers
aren't contributing to high revenue.","metadata":{}},{"cell_type":"markdown","source":"#
Question 3\n\nAre they also profitable - what is the profit margin across the
buckets","metadata":{}},{"cell_type":"code","source":"profit_df =
pd.DataFrame(columns=tmp_df.columns)","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.793411Z","iopub.status.idle":"2022-11-
20T06:55:03.794401Z","shell.execute_reply.started":"2022-11-
20T06:55:03.794068Z","shell.execute_reply":"2022-11-
20T06:55:03.794100Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"for g, d in grouped_object:\n   highest_profit =
d.nlargest(1, 'Profit')\n   profit_df = pd.concat([profit_df,highest_profit])","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.796181Z","iopub.status.idle":"2022-
11-20T06:55:03.797134Z","shell.execute_reply.started":"2022-11-
20T06:55:03.796823Z","shell.execute_reply":"2022-11-
20T06:55:03.796853Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"def highlight_cols(x): \n   df = x.copy()\n   df.loc[:,
['Customer Name','Profit']] = 'background-color: green'\n
df[['Order_year','Purchase_during_year','unique_customers_within_year','Revenue','Customer_p
urchase_frequency']] = 'background-color: grey'\n   return df ","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.799053Z","iopub.status.idle":"2022-11-
20T06:55:03.800016Z","shell.execute_reply.started":"2022-11-

20T06:55:03.799707Z","shell.execute_reply":"2022-11-
20T06:55:03.799737Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"display(profit_df.style.apply(highlight_cols, axis =
None))","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.801818Z","iopub.status.idle":"2022-11-
20T06:55:03.802797Z","shell.execute_reply.started":"2022-11-
20T06:55:03.802479Z","shell.execute_reply":"2022-11-
20T06:55:03.802519Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer:\n\nFrom the table above we can see that neither
of the customers who were in high purchase frequency table or high revenue table are here in
the high profitable customer table. May be these are the customer who are purchasing low
quantity but the profit margin is higher on their purchase.","metadata":{}},
{"cell_type":"markdown","source":"# Question 4:\n\nWhich customer segment is most profitable
in each year.","metadata":{}},{"cell_type":"code","source":"segment_group =
data.groupby(['Order_year','Segment'])","metadata":{"execution":{"iopub.status.busy":"2022-
11-20T06:55:03.804583Z","iopub.status.idle":"2022-11-
20T06:55:03.805563Z","shell.execute_reply.started":"2022-11-
20T06:55:03.805219Z","shell.execute_reply":"2022-11-
20T06:55:03.805251Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"high_profit_df =
segment_group.agg({'Profit':'sum'}).unstack()\nhigh_profit_df.style.background_gradient(cmap
='Spectral', subset=pd.IndexSlice[:, pd.IndexSlice[:,'Consumer']])","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.807304Z","iopub.status.idle":"2022-11-
20T06:55:03.808275Z","shell.execute_reply.started":"2022-11-
20T06:55:03.807974Z","shell.execute_reply":"2022-11-
20T06:55:03.808004Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer:\n\nWe can see that every year consumer segment
is triggering more profit to the firm.","metadata":{}},{"cell_type":"markdown","source":"#
Question: 4\n\nHow the customers are distributed across the countries?","metadata":{}},
{"cell_type":"code","source":"country_group = data.groupby(['Country'])","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.810074Z","iopub.status.idle":"2022-
11-20T06:55:03.811020Z","shell.execute_reply.started":"2022-11-
20T06:55:03.810710Z","shell.execute_reply":"2022-11-
20T06:55:03.810740Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"customer_distribution = country_group.agg({'Customer
ID':'count'})\ncustomer_distribution.columns =
['Customer_count']\ncustomer_distribution.reset_index(inplace=True)\ncustomer_distribution",
"metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.812376Z","iopub.status.idle":"2022-11-
20T06:55:03.813011Z","shell.execute_reply.started":"2022-11-
20T06:55:03.812808Z","shell.execute_reply":"2022-11-
20T06:55:03.812829Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"country_map = dict(type='choropleth',\n
locations=customer_distribution['Country'],\n   locationmode='country names',\n
z=customer_distribution['Customer_count'],\n   reversescale = True,\n
text=customer_distribution['Country'],\n   colorscale='earth',\n   colorbar=
{'title':'Customer Count'})","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.814120Z","iopub.status.idle":"2022-11-
20T06:55:03.863063Z","shell.execute_reply.started":"2022-11-
20T06:55:03.862570Z","shell.execute_reply":"2022-11-
20T06:55:03.862616Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"layout = dict(title='Customer Distribution over Countries',\n
geo=dict(showframe=False,projection={'type':'mercator'}))","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.865139Z","iopub.status.idle":"2022-11-
20T06:55:03.866151Z","shell.execute_reply.started":"2022-11-
20T06:55:03.865835Z","shell.execute_reply":"2022-11-
20T06:55:03.865869Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer: ","metadata":{}},
{"cell_type":"code","source":"choromap = go.Figure(data = [country_map],layout =
layout)\niplot(choromap)","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.867902Z","iopub.status.idle":"2022-11-
20T06:55:03.868899Z","shell.execute_reply.started":"2022-11-
20T06:55:03.868537Z","shell.execute_reply":"2022-11-
20T06:55:03.868567Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Product Analysis\n\n**Question: 1**Which country has top

sales?\n\n**Question: 2**Which are the top 5 profit-making product types on a yearly basis\n\n**Question: 3**How is the product price varying with sales - Is there any increase in sales with the decrease in price at a day level\n\n**Question: 4**What is the average delivery time across the counties - bar plot","metadata":{}},
{"cell_type":"markdown","source":"# Question: 1\n\nWhich country has top sales?","metadata":{}},{"cell_type":"code","source":"country_group = data.groupby('Country')","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.870594Z","iopub.status.idle":"2022-11-20T06:55:03.871542Z","shell.execute_reply.started":"2022-11-20T06:55:03.871206Z","shell.execute_reply":"2022-11-20T06:55:03.871236Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"country_sales = country_group.agg({'Sales':'sum'})\ncountry_sales.sort_values(by='Sales', ascending=False)","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.873245Z","iopub.status.idle":"2022-11-20T06:55:03.874195Z","shell.execute_reply.started":"2022-11-20T06:55:03.873886Z","shell.execute_reply":"2022-11-20T06:55:03.873916Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer: \n\n\nWe can see that United States has top sales. Things are better when they are visually presented. Let's plot top 10 sales countries.","metadata":{}},{"cell_type":"code","source":"import squarify","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.875890Z","iopub.status.idle":"2022-11-20T06:55:03.876816Z","shell.execute_reply.started":"2022-11-20T06:55:03.876511Z","shell.execute_reply":"2022-11-20T06:55:03.876541Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"top_10_sales = country_sales.nlargest(10, 'Sales')\ntop_10_sales.index","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.878495Z","iopub.status.idle":"2022-11-20T06:55:03.879550Z","shell.execute_reply.started":"2022-11-20T06:55:03.879212Z","shell.execute_reply":"2022-11-20T06:55:03.879242Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"plt.figure(figsize=(15,7))\nrevs = top_10_sales['Sales'].values\nlabels = ['United States: 2297200.8603',\n   'Australia: 925235.853',\n   'France: 858931.083',\n   'China: 700562.025',\n   'Germany: 628840.0305',\n   'Mexico: 622590.61752',\n   'India: 589650.105',\n   'United Kingdom: 528576.3',\n   'Indonesia: 404887.4979',\n   'Brazil: 361106.41896']\nsquarify.plot(revs, label=labels,color= sns.color_palette('copper'), alpha=0.7)\nplt.show()","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.881232Z","iopub.status.idle":"2022-11-20T06:55:03.882186Z","shell.execute_reply.started":"2022-11-20T06:55:03.881877Z","shell.execute_reply":"2022-11-20T06:55:03.881907Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Question 2:\n\nWhich are the top 5 profit-making product types on a yearly basis","metadata":{}},{"cell_type":"code","source":"year_category_group = data.groupby(['Order_year','Sub-Category'])","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.884096Z","iopub.status.idle":"2022-11-20T06:55:03.885028Z","shell.execute_reply.started":"2022-11-20T06:55:03.884717Z","shell.execute_reply":"2022-11-20T06:55:03.884747Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"year_category_proft_df = year_category_group.agg({'Profit':'sum'})\nyear_category_proft_df","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.886739Z","iopub.status.idle":"2022-11-20T06:55:03.887683Z","shell.execute_reply.started":"2022-11-20T06:55:03.887365Z","shell.execute_reply":"2022-11-20T06:55:03.887397Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"year_category_proft_df.reset_index(inplace=True)\ncategory_yearly_profit = year_category_proft_df.groupby('Order_year')\ntop5_profit_category = pd.DataFrame(columns=year_category_proft_df.columns)","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.889380Z","iopub.status.idle":"2022-11-20T06:55:03.890292Z","shell.execute_reply.started":"2022-11-20T06:55:03.889983Z","shell.execute_reply":"2022-11-20T06:55:03.890012Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"for g, d in category_yearly_profit:\n   high_profit_categories = d.nlargest(5, 'Profit')\n   top5_profit_category = pd.concat([top5_profit_category,high_profit_categories])","metadata":{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.891981Z","iopub.status.idle":"2022-11-

20T06:55:03.892928Z","shell.execute_reply.started":"2022-11-
20T06:55:03.892624Z","shell.execute_reply":"2022-11-
20T06:55:03.892655Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer: \n\nBelow dataframe includes top 5 profit making
products for each year.","metadata":{}},
{"cell_type":"code","source":"top5_profit_category.style.background_gradient(cmap='Spectral'
, subset=pd.IndexSlice[:, 'Profit'])","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.894802Z","iopub.status.idle":"2022-11-
20T06:55:03.895839Z","shell.execute_reply.started":"2022-11-
20T06:55:03.895431Z","shell.execute_reply":"2022-11-
20T06:55:03.895464Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Question 3:\n\nHow is the product price varying with
sales - Is there any increase in sales with the decrease in price at a day
level","metadata":{}},{"cell_type":"markdown","source":"**Note**: This question could have
been more specifir to a coutry/product category and for a specific year but all we have
asked is to see the trend between sales and price. Let's try to plot it as asked.\n\nHow to
Calculate Unit Price. The unit price can be found using a simple formula if the quantity and
total cost is known. Simply divide the total price by the quantity to find the unit price.
Thus, the unit price is equal to the total price divided by the quantity.","metadata":{}},
{"cell_type":"code","source":"data['Unit_price'] =
data['Sales']/data['Quantity']\ndata['Order_day'] = data['Order Date'].dt.day","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.897670Z","iopub.status.idle":"2022-
11-20T06:55:03.898609Z","shell.execute_reply.started":"2022-11-
20T06:55:03.898264Z","shell.execute_reply":"2022-11-
20T06:55:03.898294Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer:\n\nFrom below chart we can see that when the
prices are lower sales are high and when the prices increase, sales decrease.","metadata":
{}},{"cell_type":"code","source":"g5 = sns.FacetGrid(data, row = 'Order_day', col =
'Order_year', hue = 'Order_day')\nkwe = dict(s = 50, linewidth = 0.5, edgecolor =
'black')\ng5 = g5.map(plt.scatter, 'Unit_price', 'Sales')\ng5.set(xlim=(0,100), ylim=
(0,100))\nfor ax in g5.axes.flat:\n   ax.plot((0,100),(0,100), c = 'gray', ls = '--
')\ng5.add_legend()","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.900299Z","iopub.status.idle":"2022-11-
20T06:55:03.901225Z","shell.execute_reply.started":"2022-11-
20T06:55:03.900917Z","shell.execute_reply":"2022-11-
20T06:55:03.900946Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Question: 4\n\nWhat is the average delivery time across
the counties - bar plot","metadata":{}},
{"cell_type":"code","source":"data['Delivery_duration'] = data['Ship Date']-data['Order
Date']\ncountry_group = data.groupby('Country')\ndelivery_duration_df =
country_group.agg({'Delivery_duration':'mean'})\ndelivery_duration_df['Duration_in_hours'] =
delivery_duration_df['Delivery_duration'] / dt.timedelta(hours=1)","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.902942Z","iopub.status.idle":"2022-11-
20T06:55:03.903861Z","shell.execute_reply.started":"2022-11-
20T06:55:03.903558Z","shell.execute_reply":"2022-11-
20T06:55:03.903589Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"We have 147 unique countries in our dataframe. Plotting
them altogether on a same plot wouln't be possible. So I will only plot top 10 sales
countries.","metadata":{}},{"cell_type":"code","source":"delivery_duration_df","metadata":
{"execution":{"iopub.status.busy":"2022-11-20T06:55:03.905530Z","iopub.status.idle":"2022-
11-20T06:55:03.906447Z","shell.execute_reply.started":"2022-11-
20T06:55:03.906120Z","shell.execute_reply":"2022-11-
20T06:55:03.906149Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"markdown","source":"# Answer:","metadata":{}},
{"cell_type":"code","source":"top10_sales_country_DD
=top_10_sales.merge(delivery_duration_df, how='left', left_index=True,
right_index=True)\ntop10_sales_country_DD.reset_index(inplace=True)\ntop10_sales_country_DD.
sort_values(by='Duration_in_hours')\n# we can see that China and Brazil are providing
fastest deliveries","metadata":{"execution":{"iopub.status.busy":"2022-11-
20T06:55:03.908111Z","iopub.status.idle":"2022-11-
20T06:55:03.909035Z","shell.execute_reply.started":"2022-11-
20T06:55:03.908729Z","shell.execute_reply":"2022-11-
20T06:55:03.908758Z"},"trusted":true},"execution_count":null,"outputs":[]},
{"cell_type":"code","source":"top10_sales_country_DD.iplot(kind='bar',x='Country',
y='Duration_in_hours',\n   title= 'Countries And their Average

Product Delivery Duration in Hours',\n   xTitle='Countries',\n
yTitle= 'AVG Delivery Duration in hours')","metadata":{"execution":
{"iopub.status.busy":"2022-11-20T06:55:03.910484Z","iopub.status.idle":"2022-11-
20T06:55:03.911123Z","shell.execute_reply.started":"2022-11-
20T06:55:03.910913Z","shell.execute_reply":"2022-11-
20T06:55:03.910934Z"},"trusted":true},"execution_count":null,"outputs":[]}]}