```
************************************************************;
************** CREATING a NEW LIBRARY FOR DATA **************;
************************************************************;
libname A4 "/folders/myfolders/DANA/Assignment 4/A4_Library";

************************************************************;
*******************IMPORTING YOUTUBE DATA*******************;
************************************************************;
PROC IMPORT
    DATAFILE='/folders/myfolders/DANA/Assignment 4/online_retail_II.xlsx'
    DBMS=XLSX
    OUT=A4.OnlineRetail REPLACE ;
    SHEET='Year 2010-2011';
RUN;


************************************************************;
********************* ANALYSING DATASETS **********************;
************************************************************;

*---------------------Analysing Table Attributes---------------;

ods noproctitle;
ods select attributes position;

title 'Data Structure for Online Retail Store';
proc contents data=A4.onlineretail order=varnum;
run;


*----------------------Analysing First 15 Rows-----------------;

proc print data=A4.onlineretail (obs=15) obs="S.No." label n;
    title 'List Data for Online Retail Transactions';
run;

*------------------Analyze numerical variables-----------------;

proc means data=A4.onlineretail n nmiss min mean median max std;
    var Price Quantity;
    title "Descriptive Statistics for Numeric Variables";
run;

*-----------------Analyze categorical variables----------------;

proc freq data=A4.onlineretail;
    tables Country/ nocum ;
    title "Frequencies for Categorical Variable Country";
run;


*-----------------Analysing Missing Values--------------------;

ods noproctitle;

proc format;
    value _nmissprint low-high="Non-missing";
    value $_cmissprint " "=" " other="Non-missing";
run;

proc freq data=A4.ONLINERETAIL;
    title3 "Missing Data Frequencies";
    title4 h=2 "Legend: ., A, B, etc = Missing";
```

```
    format InvoiceDate Customer_ID _nmissprint.;
    format Invoice StockCode Description Country $_cmissprint.;
    tables Invoice StockCode Description InvoiceDate Customer_ID
        Country / missing nocum;
run;

proc freq data=A4.ONLINERETAIL noprint;
    table Invoice * StockCode * Description * InvoiceDate *
        Customer_ID * Country / missing out=Work._MissingData_;
    format InvoiceDate Customer_ID _nmissprint.;
    format Invoice StockCode Description Country $_cmissprint.;
run;

proc print data=Work._MissingData_ noobs label;
    title3 "Missing Data Patterns across Variables";
    title4 h=2 "Legend: ., A, B, etc = Missing";
    format InvoiceDate Customer_ID _nmissprint.;
    format Invoice StockCode Description Country $_cmissprint.;
    label count="Frequency" percent="Percent";
run;

title3;

/* Clean up */
proc delete data=Work._MissingData_;
run;

*--------------Analyze Canacelled and Invalid Orders-----------;

proc sql;
title "Number of Orders Marked Cancelled";
select COUNT(*) as Canceled_Orders
from A4.OnlineRetail where Invoice Like 'C%';
run;

proc sql;
title "Number of Orders Marked Cancelled with Negative Quantities";
select COUNT(*) as Canceled_Orders
from A4.OnlineRetail where Invoice Like 'C%' and Quantity < 0;
run;

proc sql;
title "Number of Orders with Negative Quntities";
select COUNT(*) as Invalid_Quantity
from A4.OnlineRetail where Quantity < 0;
run;

proc sql;
title "Number of Orders with Negative Price";
select COUNT(*) as Invalid_Price
from A4.OnlineRetail where Price < 0;
run;


************************************************************;
**********************Cleaning DATASETS**********************;
************************************************************;

*-------------------Removing Invalid Records-----------------;

data WORK.CANCELLED_REMOVED;
    set A4.onlineretail;
    where Invoice not like 'C%';
    CustomerID = PUT(Customer_ID, $12.);
```

```
    drop Customer_ID;
run;


data WORK.Positive_QTY;
    set WORK.CANCELLED_REMOVED;
    where Quantity>=0;
    rename CustomerID=Customer_ID;
run;

data WORK.Positive_Price;
    set WORK.Positive_QTY;
    where Price>=0;
run;

*-----------------Handling Missing Values----------------------;

data A4.OnlineRetail_Cleaned;
    set WORK.Positive_Price;
    if Description='' then Description='N/A';
    if strip(Customer_ID)='.' then Customer_ID='UNREGISTERED';
    else Customer_ID=strip(Customer_ID);
run;


************************************************************;
**********Analysing Cleaned DATASETS**************;
************************************************************;


*-------------------- Analysing Table Attributes --------------;

ods noproctitle;
ods select attributes position;

title 'Data Structure for Online Retail Store';
proc contents data=A4.OnlineRetail_Cleaned order=varnum;
run;


*-----------------Analysing First 15 Rows----------------------;

proc print data=A4.OnlineRetail_Cleaned (obs=15) obs="S.No." label n;
    title 'List Data for Online Retail Transactions';
run;

*------------------Analyze numerical variables-----------------;

proc means data=A4.OnlineRetail_Cleaned n nmiss min mean median max std;
    var Price Quantity;
    title "Descriptive Statistics for Numeric Variables";
run;

*-----------------Analyze categorical variables----------------;

proc freq data=A4.OnlineRetail_Cleaned;
    tables Country/ nocum ;
    title "Frequencies for Categorical Variable Country";
run;


*-----------------Analysing Missing Values--------------------;

ods noproctitle;
```

```
192
193 proc format;
194     value _nmissprint low-high="Non-missing";
195     value $_cmissprint " "=" " other="Non-missing";
196 run;
197
198 proc freq data=A4.ONLINERETAIL_CLEANED;
199     title3 "Missing Data Frequencies";
200     title4 h=2 "Legend: ., A, B, etc = Missing";
201     format Quantity InvoiceDate Price _nmissprint.;
202     format Invoice StockCode Description Country
203            Customer_ID $_cmissprint.;
204     tables Invoice StockCode Description Quantity
205            InvoiceDate Price Country
206         Customer_ID / missing nocum;
207 run;
208
209 proc freq data=A4.ONLINERETAIL_CLEANED noprint;
210     table Invoice * StockCode * Description * Quantity *
211           InvoiceDate * Price * Country * Customer_ID / missing
212           out=Work._MissingData_;
213     format Quantity InvoiceDate Price _nmissprint.;
214     format Invoice StockCode Description Country
215            Customer_ID $_cmissprint.;
216 run;
217
218 proc print data=Work._MissingData_ noobs label;
219     title3 "Missing Data Patterns across Variables";
220     title4 h=2 "Legend: ., A, B, etc = Missing";
221     format Quantity InvoiceDate Price _nmissprint.;
222     format Invoice StockCode Description Country
223            Customer_ID $_cmissprint.;
224     label count="Frequency" percent="Percent";
225 run;
226
227 title3;
228
229 /* Clean up */
230 proc delete data=Work._MissingData_;
231 run;
232
233
234 proc sql;
235 title "Number of Orders with Unregistered Customers";
236 select COUNT(*) as Unregistered_Customers
237 from A4.OnlineRetail_Cleaned where Customer_ID = 'UNREGISTERED';
238 run;
239
240 proc sql;
241 title "Number of Orders with No Description";
242 select COUNT(*) as No_Description
243 from A4.OnlineRetail_Cleaned where Description = 'N/A';
244 run;
245
246 *------------------------daily unit sales----------------------;
247 data Work.DailyUnits;
248 set A4.OnlineRetail_Cleaned;
249 date_new = datepart(InvoiceDate);
250 format date_new date9.;
251 keep date_new Quantity;
252 run;
253
254 proc sql;
255 create table Work.DailyUnitSales as
```

```
select date_new, sum(Quantity) as Units_Sold from Work.DailyUnits group by date_new;
run;

*----------------Unusual Daily unit sales----------------------;

ods graphics / reset width=12in height=6in imagemap;

proc sort data=WORK.DAILYUNITSALES out=_SeriesPlotTaskData;
    by date_new;
run;

proc sgplot data=_SeriesPlotTaskData;
    title height=14pt "Daily Unit Sales";
    series x=date_new y=Units_Sold / markers markerattrs=(symbol=circlefilled);
    xaxis grid label="Invoice Date" valuesrotate=diagonal;
    yaxis grid label="Number of Units Sold";
run;

ods graphics / reset;
title;

*------------------Time Lag in Units Sold----------------------;

data Work.Unit_Sales_By_Day_Hour;
set A4.OnlineRetail_Cleaned;
Hour = Hour(InvoiceDate);
Day = datepart(InvoiceDate);
format Day DOWNAME.;
keep Hour Day Quantity;
run;

ods graphics / reset width=8in height=4.8in imagemap;

proc sgplot data=WORK.UNIT_SALES_BY_DAY_HOUR;
    title height=14pt "Hourly Unit Sales";
    vbar Hour / response=Quantity datalabel;
    yaxis;
    xaxis grid;
run;

ods graphics / reset;
title;

ods graphics / reset width=6.5in height=4.8in imagemap;

proc sgplot data=WORK.UNIT_SALES_BY_DAY_HOUR;
    title height=14pt "Weekly Unit Sales";
    vbar Day / response=Quantity datalabel;
    yaxis grid;
run;

ods graphics / reset;
title;

*--------------Monetary Sales in Thousands Per Month-----------;

data Work.OnlineSales;
set A4.OnlineRetail_Cleaned;
Month = datepart(InvoiceDate);
Sales = Price*Quantity/1000;
format Month MONYY5.;
keep Month Sales;
run;

```

```
ods graphics / reset width=6.5in height=4.8in imagemap;

proc sgplot data=WORK.ONLINESALES;
    title height=14pt "Monthly Monetary Sales (in Thousands)";
    vbar Month / response=Sales datalabel;
    yaxis grid label="Sales (Sum) in thousands";
run;

ods graphics / reset;
title;

*----------Product Contribution to the Monetary Sales----------;

proc sql outobs=10;
create table Work.SalesByProduct as
select StockCode, SUM(Price*Quantity) as Sales
from A4.OnlineRetail_Cleaned
group by StockCode
order by 2 desc;
run;

proc sql outobs=1;
select a.StockCode, a.Description, b.Sales
from A4.OnlineRetail_Cleaned a, Work.SalesByProduct b
where a.StockCode=b.StockCode
order by 3 desc;
run;

ods graphics / reset width=6.5in height=4.8in imagemap discretemax=4000;

proc sgplot data=Work.SalesByProduct;
    title height=14pt "Monetary Sales by Product";
    vbar StockCode / response=Sales datalabel;
    yaxis grid label="Sales (Sum) in thousands" discreteorder=data;
run;

ods graphics / reset;
title;

*----------------------Sales Projection-----------------------;

Proc Sql;
Select Sum(Price*Quantity) as Total_Sales from A4.onlineretail_cleaned;
Quit;

Proc Sql noprint;
Select Sum(Price*Quantity) into :total_Sales
from A4.onlineretail_cleaned;
Quit;
%Put &total_Sales;*10666703;


proc sql;
create table Work.SalesByProduct as
select StockCode, SUM(Price*Quantity) as Sales
from A4.OnlineRetail_Cleaned
group by StockCode
order by 2;
run;


data work.CumulativeSales;
set Work.SalesByProduct;
retain sales_cum;
```

```
if _n_ = 1 then sales_cum=sales;
else sales_cum=sales_cum+sales;
run;

/* proc sql; */
/* select * from work.CumulativeSales where sales_cum<=&total_Sales*0.2; */
/* run; */

proc sql;
create table Work.Top80Sales as
select * from A4.OnlineRetail_Cleaned
where StockCode not in (select StockCode from work.CumulativeSales
                        where sales_cum<=&total_Sales*0.2);
run;

/* Proc Sql noprint; */
/* Select Sum(Price*Quantity) into :total_Sales */
/* from work.Top80Sales; */
/* Quit; */
/* %Put &total_Sales;*8534481; */

proc sql outobs=10;
create table Work.Top10Sales as
select StockCode from work.CumulativeSales
order by Sales desc;
run;

/* Proc Sql noprint; */
/* Select Sum(Price*Quantity) into :total_Sales */
/* from Work.Top80Sales where StockCode in (Select StockCode from Work.Top10Sales); */
/* Quit; */
/* %Put &total_Sales;*1152462; */

proc sql;
create table Work.TargetSales as
select Invoice, StockCode, Description,CEIL(Quantity*1.1) as Quantity,
       InvoiceDate, Price, Customer_ID, Country
from Work.Top80Sales where StockCode in (select StockCode from Work.Top10Sales);
run;

proc sql;
create table Work.OtherSales as
select * from Work.Top80Sales
where StockCode not in (select StockCode from Work.Top10Sales);
run;

data work.ProjectedSales;
set WORK.TARGETSALES WORK.OTHERSALES;
Sales = Price*Quantity;
run;

/* Proc Sql noprint; */
/* Select Sum(Sales) into :total_Sales */
/* from work.ProjectedSales; */
/* Quit; */
/* %Put &total_Sales;*8963175; */


proc sort data=Work.PROJECTEDSALES out=A4.ProjectedSales;
    by InvoiceDate;
run;

*-----------------Projected Sales by Month--------------------;

```

```
448
449 data Work.OnlineSalesProj;
450 set A4.ProjectedSales;
451 Month = datepart(InvoiceDate);
452 Sales_Th = Sales/1000;
453 format Month MONYY5.;
454 keep Month Sales_Th;
455 run;
456
457 ods graphics / reset width=6.5in height=4.8in imagemap;
458
459 proc sgplot data=WORK.ONLINESALESProj;
460     title height=14pt "Projected Monthly Monetary Sales (in Thousands)";
461     vbar Month / response=Sales_Th datalabel;
462     yaxis grid label="Sales (Sum) in thousands";
463 run;
464
465 ods graphics / reset;
466 title;
467
468
469 *------------------Projected Total Sale-----------------------;
470 Proc Sql;
471 Select Sum(Sales) as Projected_Total_Sales from A4.ProjectedSales;
472 Quit;
473
474
475 ************************************************************;
476 **********Exporting Cleaned DATASETS**************;
477 ************************************************************;
478 proc export data=A4.OnlineRetail_Cleaned
479     outfile="/folders/myfolders/DANA/Assignment 4/CLEANED_ONLINE_RETAIL.xlsx"
480     dbms=xlsx
481     replace;
482 run;
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
```