

```

1 *****;
2 ***** CREATING a NEW LIBRARY FOR DATA *****;
3 *****;
4 libname A4 "/folders/myfolders/DANA/Assignment 4/A4_Library";
5
6 *****;
7 *****IMPORTING YOUTUBE DATA*****;
8 *****;
9 PROC IMPORT
10     DATAFILE='/folders/myfolders/DANA/Assignment 4/online_retail_II.xlsx'
11     DBMS=XLSX
12     OUT=A4.OnlineRetail REPLACE ;
13     SHEET='Year 2010-2011';
14 RUN;
15
16
17 *****;
18 ***** ANALYSING DATASETS *****;
19 *****;
20
21 *-----Analysing Table Attributes-----;
22
23 ods noproctitle;
24 ods select attributes position;
25
26 title 'Data Structure for Online Retail Store';
27 proc contents data=A4.onlineretail order=varnum;
28 run;
29
30
31 *-----Analysing First 15 Rows-----;
32
33 proc print data=A4.onlineretail (obs=15) obs="S.No." label n;
34     title 'List Data for Online Retail Transactions';
35 run;
36
37 *-----Analyze numerical variables-----;
38
39 proc means data=A4.onlineretail n nmiss min mean median max std;
40     var Price Quantity;
41     title "Descriptive Statistics for Numeric Variables";
42 run;
43
44 *-----Analyze categorical variables-----;
45
46 proc freq data=A4.onlineretail;
47     tables Country/ nocum ;
48     title "Frequencies for Categorical Variable Country";
49 run;
50
51
52 *-----Analysing Missing Values-----;
53
54 ods noproctitle;
55
56 proc format;
57     value _nmissprint low-high="Non-missing";
58     value $_cmissprint " "=" " other="Non-missing";
59 run;
60
61 proc freq data=A4.ONLINERETAIL;
62     title3 "Missing Data Frequencies";
63     title4 h=2 "Legend: ., A, B, etc = Missing";

```

```

64     format InvoiceDate Customer_ID _nmissprint.;
65     format Invoice StockCode Description Country $_cmissprint.;
66     tables Invoice StockCode Description InvoiceDate Customer_ID
67            Country / missing nocum;
68 run;
69
70 proc freq data=A4.ONLINERETAIL noprint;
71     table Invoice * StockCode * Description * InvoiceDate *
72            Customer_ID * Country / missing out=Work._MissingData_;
73     format InvoiceDate Customer_ID _nmissprint.;
74     format Invoice StockCode Description Country $_cmissprint.;
75 run;
76
77 proc print data=Work._MissingData_ noobs label;
78     title3 "Missing Data Patterns across Variables";
79     title4 h=2 "Legend: ., A, B, etc = Missing";
80     format InvoiceDate Customer_ID _nmissprint.;
81     format Invoice StockCode Description Country $_cmissprint.;
82     label count="Frequency" percent="Percent";
83 run;
84
85 title3;
86
87 /* Clean up */
88 proc delete data=Work._MissingData_;
89 run;
90
91 *-----Analyze Canacelled and Invalid Orders-----;
92
93 proc sql;
94     title "Number of Orders Marked Cancelled";
95     select COUNT(*) as Canceled_Orders
96     from A4.OnlineRetail where Invoice Like 'C%';
97 run;
98
99 proc sql;
100    title "Number of Orders Marked Cancelled with Negative Quantities";
101    select COUNT(*) as Canceled_Orders
102    from A4.OnlineRetail where Invoice Like 'C%' and Quantity < 0;
103 run;
104
105 proc sql;
106    title "Number of Orders with Negative Quntities";
107    select COUNT(*) as Invalid_Quantity
108    from A4.OnlineRetail where Quantity < 0;
109 run;
110
111 proc sql;
112    title "Number of Orders with Negative Price";
113    select COUNT(*) as Invalid_Price
114    from A4.OnlineRetail where Price < 0;
115 run;
116
117
118 *****;
119 *****Cleaning DATASETS*****;
120 *****;
121
122 *-----Removing Invalid Records-----;
123
124 data WORK.CANCELLED_REMOVED;
125     set A4.onlineretail;
126     where Invoice not like 'C%';
127     CustomerID = PUT(Customer_ID, $12.);

```

```

128     drop Customer_ID;
129 run;
130
131
132 data WORK.Positive_QTY;
133     set WORK.CANCELLED_REMOVED;
134     where Quantity>=0;
135     rename CustomerID=Customer_ID;
136 run;
137
138 data WORK.Positive_Price;
139     set WORK.Positive_QTY;
140     where Price>=0;
141 run;
142
143 *-----Handling Missing Values-----;
144
145 data A4.OnlineRetail_Cleaned;
146     set WORK.Positive_Price;
147     if Description='' then Description='N/A';
148     if strip(Customer_ID)='.' then Customer_ID='UNREGISTERED';
149     else Customer_ID=strip(Customer_ID);
150 run;
151
152
153 *****;
154 *****Analysing Cleaned DATASETS*****;
155 *****;
156
157
158 *----- Analysing Table Attributes -----;
159
160 ods noproctitle;
161 ods select attributes position;
162
163 title 'Data Structure for Online Retail Store';
164 proc contents data=A4.OnlineRetail_Cleaned order=varnum;
165 run;
166
167
168 *-----Analysing First 15 Rows-----;
169
170 proc print data=A4.OnlineRetail_Cleaned (obs=15) obs="S.No." label n;
171     title 'List Data for Online Retail Transactions';
172 run;
173
174 *-----Analyze numerical variables-----;
175
176 proc means data=A4.OnlineRetail_Cleaned n nmiss min mean median max std;
177     var Price Quantity;
178     title "Descriptive Statistics for Numeric Variables";
179 run;
180
181 *-----Analyze categorical variables-----;
182
183 proc freq data=A4.OnlineRetail_Cleaned;
184     tables Country/ nocum ;
185     title "Frequencies for Categorical Variable Country";
186 run;
187
188
189 *-----Analysing Missing Values-----;
190
191 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

```

```

256 select date_new, sum(Quantity) as Units_Sold from Work.DailyUnits group by date_new;
257 run;
258
259 *-----Unusual Daily unit sales-----;
260
261 ods graphics / reset width=12in height=6in imagemap;
262
263 proc sort data=WORK.DAILYUNITSALES out=_SeriesPlotTaskData;
264     by date_new;
265 run;
266
267 proc sgplot data=_SeriesPlotTaskData;
268     title height=14pt "Daily Unit Sales";
269     series x=date_new y=Units_Sold / markers markerattrs=(symbol=circlefilled);
270     xaxis grid label="Invoice Date" valuesrotate=diagonal;
271     yaxis grid label="Number of Units Sold";
272 run;
273
274 ods graphics / reset;
275 title;
276
277 *-----Time Lag in Units Sold-----;
278
279 data Work.Unit_Sales_By_Day_Hour;
280 set A4.OnlineRetail_Cleaned;
281 Hour = Hour(InvoiceDate);
282 Day = datepart(InvoiceDate);
283 format Day DOWNNAME.;
284 keep Hour Day Quantity;
285 run;
286
287 ods graphics / reset width=8in height=4.8in imagemap;
288
289 proc sgplot data=WORK.UNIT_SALES_BY_DAY_HOUR;
290     title height=14pt "Hourly Unit Sales";
291     vbar Hour / response=Quantity datalabel;
292     yaxis;
293     xaxis grid;
294 run;
295
296 ods graphics / reset;
297 title;
298
299 ods graphics / reset width=6.5in height=4.8in imagemap;
300
301 proc sgplot data=WORK.UNIT_SALES_BY_DAY_HOUR;
302     title height=14pt "Weekly Unit Sales";
303     vbar Day / response=Quantity datalabel;
304     yaxis grid;
305 run;
306
307 ods graphics / reset;
308 title;
309
310 *-----Monetary Sales in Thousands Per Month-----;
311
312 data Work.OnlineSales;
313 set A4.OnlineRetail_Cleaned;
314 Month = datepart(InvoiceDate);
315 Sales = Price*Quantity/1000;
316 format Month MONYY5.;
317 keep Month Sales;
318 run;
319

```

```

320 ods graphics / reset width=6.5in height=4.8in imagemap;
321
322 proc sgplot data=WORK.ONLINESALES;
323     title height=14pt "Monthly Monetary Sales (in Thousands)";
324     vbar Month / response=Sales datalabel;
325     yaxis grid label="Sales (Sum) in thousands";
326 run;
327
328 ods graphics / reset;
329 title;
330
331 *-----Product Contribution to the Monetary Sales-----;
332
333 proc sql outobs=10;
334 create table Work.SalesByProduct as
335 select StockCode, SUM(Price*Quantity) as Sales
336 from A4.OnlineRetail_Cleaned
337 group by StockCode
338 order by 2 desc;
339 run;
340
341 proc sql outobs=1;
342 select a.StockCode, a.Description, b.Sales
343 from A4.OnlineRetail_Cleaned a, Work.SalesByProduct b
344 where a.StockCode=b.StockCode
345 order by 3 desc;
346 run;
347
348 ods graphics / reset width=6.5in height=4.8in imagemap discretemax=4000;
349
350 proc sgplot data=Work.SalesByProduct;
351     title height=14pt "Monetary Sales by Product";
352     vbar StockCode / response=Sales datalabel;
353     yaxis grid label="Sales (Sum) in thousands" discreteorder=data;
354 run;
355
356 ods graphics / reset;
357 title;
358
359 *-----Sales Projection-----;
360
361 Proc Sql;
362 Select Sum(Price*Quantity) as Total_Sales from A4.onlineretail_cleaned;
363 Quit;
364
365 Proc Sql noprint;
366 Select Sum(Price*Quantity) into :total_Sales
367 from A4.onlineretail_cleaned;
368 Quit;
369 %Put &total_Sales;*10666703;
370
371
372 proc sql;
373 create table Work.SalesByProduct as
374 select StockCode, SUM(Price*Quantity) as Sales
375 from A4.OnlineRetail_Cleaned
376 group by StockCode
377 order by 2;
378 run;
379
380
381 data work.CumulativeSales;
382 set Work.SalesByProduct;
383 retain sales_cum;

```

```

384 if _n_ = 1 then sales_cum=sales;
385 else sales_cum=sales_cum+sales;
386 run;
387
388 /* proc sql; */
389 /* select * from work.CumulativeSales where sales_cum<=&total_Sales*0.2; */
390 /* run; */
391
392 proc sql;
393 create table Work.Top80Sales as
394 select * from A4.OnlineRetail_Cleaned
395 where StockCode not in (select StockCode from work.CumulativeSales
396                        where sales_cum<=&total_Sales*0.2);
397 run;
398
399 /* Proc Sql noprint; */
400 /* Select Sum(Price*Quantity) into :total_Sales */
401 /* from work.Top80Sales; */
402 /* Quit; */
403 /* %Put &total_Sales;*8534481; */
404
405 proc sql outobs=10;
406 create table Work.Top10Sales as
407 select StockCode from work.CumulativeSales
408 order by Sales desc;
409 run;
410
411 /* Proc Sql noprint; */
412 /* Select Sum(Price*Quantity) into :total_Sales */
413 /* from Work.Top80Sales where StockCode in (Select StockCode from Work.Top10Sales); */
414 /* Quit; */
415 /* %Put &total_Sales;*1152462; */
416
417 proc sql;
418 create table Work.TargetSales as
419 select Invoice, StockCode, Description, CEIL(Quantity*1.1) as Quantity,
420        InvoiceDate, Price, Customer_ID, Country
421 from Work.Top80Sales where StockCode in (select StockCode from Work.Top10Sales);
422 run;
423
424 proc sql;
425 create table Work.OtherSales as
426 select * from Work.Top80Sales
427 where StockCode not in (select StockCode from Work.Top10Sales);
428 run;
429
430 data work.ProjectedSales;
431 set WORK.TARGETSALES WORK.OTHERSALES;
432 Sales = Price*Quantity;
433 run;
434
435 /* Proc Sql noprint; */
436 /* Select Sum(Sales) into :total_Sales */
437 /* from work.ProjectedSales; */
438 /* Quit; */
439 /* %Put &total_Sales;*8963175; */
440
441
442 proc sort data=Work.PROJECTEDSALES out=A4.ProjectedSales;
443     by InvoiceDate;
444 run;
445
446 *-----Projected Sales by Month-----;
447

```

```

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

```