

```

/* Rename variables by replacing spaces with underscores for
easier reference */

data clean_railway;

set Assmt.sampled_railway_inconsistencies;

rename

  'Transaction ID'n      = Transaction_ID
  'Date of Purchase'n   = Date_of_Purchase
  'Time of Purchase'n   = Time_of_Purchase
  'Purchase Type'n      = Purchase_Type
  'Payment Method'n     = Payment_Method
  'Railcard'n            = Railcard
  'Ticket Class'n        = Ticket_Class
  'Ticket Type'n         = Ticket_Type
  'Price'n               = Price
  'Departure Station'n   = Departure_Station
  'Arrival Destination'n = Arrival_Destination
  'Date of Journey'n     = Date_of_Journey
  'Departure Time'n      = Departure_Time
  'Arrival Time'n        = Arrival_Time
  'Actual Arrival Time'n = Actual_Arrival_Time
  'Journey Status'n      = Journey_Status
  'Reason for Delay'n    = Reason_for_Delay
  'Refund Request'n      = Refund_Request;

run;

```

```

/*Display SAS Data types & format*/

proc contents data=clean_railway out=var_details(keep=name
type);

```

```
run;

/*Nominal and Ordinal Attributes*/

/* Purchase Type - Frequency count + bar chart */
proc freq data=clean_railway;
  tables Purchase_Type / plots=freqplot;
run;

/* Payment Method - Frequency table + bar chart */
proc freq data=clean_railway;
  tables Payment_Method / plots=freqplot;
run;

/* Railcard - Frequency count + bar chart */
proc freq data=clean_railway;
  tables Railcard / plots=freqplot;
run;

/* Ticket Class - Ordered categories */
proc freq data=clean_railway;
  tables Ticket_Class / plots=freqplot;
run;

/* Ticket Type - e.g., Advance, Anytime, Season */
proc freq data=clean_railway;
```

```

tables Ticket_Type / plots=freqplot;
run;

/* Journey Status - Delayed vs On Time */
proc freq data=clean_railway;
tables Journey_Status / plots=freqplot;
run;

/* Reason for Delay - Frequency + bar chart */
proc freq data=clean_railway;
tables Reason_for_Delay / plots=freqplot;
run;

/* Refund Request - Frequency of Yes/No */
proc freq data=clean_railway;
tables Refund_Request / plots=freqplot;
run;

/* Departure Station - Frequency + full bar chart */
proc freq data=clean_railway noprint;
tables Departure_Station / out=Freq_Departure;
run;

proc sgplot data=Freq_Departure;
vbar Departure_Station / response=Count datalabel;
xaxis display=(nolabel) fitpolicy=rotate;
yaxis label="Number of Departures";

```

```

run;

/* Arrival Destination - Frequency + full bar chart */
proc freq data=clean_railway noprint;
  tables Arrival_Destination / out=Freq_Arrival;
run;

proc sgplot data=Freq_Arrival;
  vbar Arrival_Destination / response=Count datalabel;
  xaxis display=(nolabel) fitpolicy=rotate;
  yaxis label="Number of Arrivals";
run;

/* Interval Attributes*/

/* Date of Purchase - Frequency distribution and line chart */
proc freq data=clean_railway noprint;
  tables Date_of_Purchase / out=Freq_Date_of_Purchase;
run;

proc sgplot data=Freq_Date_of_Purchase;
  series x=Date_of_Purchase y=Count;
  xaxis label="Date of Purchase";
  yaxis label="Number of Purchases";
run;

/* Time of Purchase - Convert to hour and visualize distribution */
*/

```

```

data clean_railway;
  set clean_railway;
  Hour_of_Purchase = hour(Time_of_Purchase);
run;

proc freq data=clean_railway;
  tables Hour_of_Purchase;
run;

proc sgplot data=clean_railway;
  histogram Hour_of_Purchase / binwidth=1;
  xaxis label="Hour of Purchase";
run;

/* Date of Journey - Frequency table and travel trend */
proc freq data=clean_railway noprint;
  tables Date_of_Journey / out=Freq_Date_of_Journey;
run;

proc sgplot data=Freq_Date_of_Journey;
  series x=Date_of_Journey y=Count;
  xaxis label="Date of Journey";
  yaxis label="Number of Journeys";
run;

/* Departure Time - Analyse by hour */
data clean_railway;

```

```
set clean_railway;
Hour_of_Departure = hour(Departure_Time);
run;

proc freq data=clean_railway;
tables Hour_of_Departure;
run;

proc sgplot data=clean_railway;
histogram Hour_of_Departure / binwidth=1;
xaxis label="Departure Hour";
run;

/* Arrival Time - Analyse by hour */
data clean_railway;
set clean_railway;
Hour_of_Arrival = hour(Arrival_Time);
run;

proc freq data=clean_railway;
tables Hour_of_Arrival;
run;

proc sgplot data=clean_railway;
histogram Hour_of_Arrival / binwidth=1;
xaxis label="Scheduled Arrival Hour";
run;
```

```

/* Actual Arrival Time - Analyse by hour */

data clean_railway;
  set clean_railway;
  Hour_of_Actual_Arrival = hour(Actual_Arrival_Time);
run;

proc freq data=clean_railway;
  tables Hour_of_Actual_Arrival;
run;

proc sgplot data=clean_railway;
  histogram Hour_of_Actual_Arrival / binwidth=1;
  xaxis label="Actual Arrival Hour";
run;

/*Ratio Attributes*/

/* Price - Descriptive statistics */

proc means data=clean_railway mean median min max std var n
maxdec=2;
  var Price;
run;

/* Price - Histogram to show distribution */

proc sgplot data=clean_railway;
  histogram Price;
  xaxis label="Ticket Price (GBP)";

```

```
run;

/* Price - Boxplot to identify outliers */
proc sgplot data=clean_railway;
  title "Boxplot of Ticket Prices";
  vbox Price;
  yaxis label="Ticket Price (GBP)";
run;

/*Data Quality - Missing values, inconcistencies etc*/
/* Summary of missing values for key fields */
proc freq data=clean_railway;
  tables
    Purchase_Type
    Payment_Method
    Railcard
    Ticket_Class
    Ticket_Type
    Departure_Station
    Arrival_Destination
    Journey_Status
    Reason_for_Delay
    Refund_Request
  / missing;
run;
```

```

proc means data=clean_railway nmiss;
run;

proc means data=clean_railway n nmiss;
var Actual_Arrival_Time;
run;

/* Frequency check to reveal inconsistent category values */
proc freq data=clean_railway;
tables Ticket_Type;
run;

proc freq data=clean_railway;
tables Reason_for_Delay;
run;

/* Check for inconsistent time formats*/
data time_issues;
set clean_railway;
flag_time_purchase = index(upcase(Time_of_Purchase), 'AM') or
index(upcase(Time_of_Purchase), 'PM');
flag_departure_time = index(upcase(Departure_Time), 'AM') or
index(upcase(Departure_Time), 'PM');
flag_arrival_time = index(upcase(Arrival_Time), 'AM') or
index(upcase(Arrival_Time), 'PM');
flag_actual_arrival = index(upcase(Actual_Arrival_Time), 'AM') or
index(upcase(Actual_Arrival_Time), 'PM');
run;

```

```

/* Display any flagged time format issues */

proc print data=time_issues;
  where flag_time_purchase = 1 or flag_departure_time = 1 or
flag_arrival_time = 1 or flag_actual_arrival = 1;
run;

/* Check inconsistent date formats */

data date_format_issues;
  set clean_railway;
  if index(Date_of_Purchase, ',') or index(Date_of_Purchase,
'.') or index(Date_of_Purchase, 'January') > 0 then flag_date = 1;
run;

proc print data=date_format_issues;
  where flag_date = 1;
run;

/* Departure Station - Frequency to reveal typos or
inconsistencies */

/* Get frequency of departure stations */

proc freq data=clean_railway noprint;
  tables Departure_Station / out=Freq_Departure_Station;
run;

/* Sort departure station frequencies in descending order */

proc sort data=Freq_Departure_Station;

```

```
by descending Count;
run;

proc print data=Freq_Departure_Station;
run;

/* Arrival Destination - Frequency to reveal typos or
inconsistencies */

proc freq data=clean_railway noprint;
  tables Arrival_Destination / out=Freq_Arrival_Destination;
run;

proc sort data=Freq_Arrival_Destination;
  by descending Count;
run;

proc print data=Freq_Arrival_Destination;
run;
```