ASSESSMENT 7

QUESTION 1

Using the merged credit data sets, created in session 6, generate 5-number summaries (minimum, 25th percentile, median, 75th percentile and maximum) for each numeric variable.

For age, amount of loan, duration of loan and instalment rate as a percentage of income, decide using the 5-number summary whether the variable is likely to be normally distributed and explain your reasoning. For each of these variables - what is the dispersion and what is the central tendency as measured by the 5-number summary?

ANSWER

VARIABLE	DISPERSION	CENTRAL TENDENCY	VARIABLE DISTRIBUTION
Age	Minimum – 19years Lower percentile 25% - 26.5 Upper percentile 75% - 42.0 Maximum – 74years	Median is 33.000 years	Majority of the data is distributed to the left. The min, 25th percentile and median values are closer than the 75th and maximum values. This suggests that the shape of the distribution is skewed to the right (positive skew).
Amount of loan	Minimum – 276.00 Lower percentile 25% - 1389.50 Upper percentile 75% - 4231.00 Maximum – 15945.00	Median - 2400	Majority of the data is distributed to the left. The min, 25th percentile and median values are closer than the 75th and maximum values. This suggests that the shape of the distribution is skewed to the right (positive skew).
Duration of loan	Minimum – 4 Lower percentile 25% - 12 Upper percentile 75% - 24 Maximum – 60	Median - 18	Majority of the data is distributed to the left. The 25th percentile and 75th percentile are close to the median equally. This suggests most of the distribution is close to the centre. The extremes are spread unevenly, the maximum value is farther from the median than the min value. This suggests that the presence of outliers to the right making the distribution positively skewed.
Instalment rate as a percentage of income	Minimum – 1 Lower percentile 25% - 2 Upper percentile 75% - 4 Maximum – 4	Median - 3	The distribution is probably normal.

Question 2

Univariate analysis: proc univariate

When compared with proc means, proc univariate provides many additional output tables. With the correct choice of options, proc means can also calculate these values but is best used when comparing sub-groups defined by categorical data.

Self-Assessment Question

Using proc univariate on age, amount of loan, duration of loan and instalment rate as a percentage of income, decide whether each variable is normally distributed and justify your answer. Compare the results obtained using proc means with that using proc univariate.

Hint: the purpose of this question is to compare the basic output of proc means with the basic output of proc univariate. Please do not use other features until asked to do so.

ANSWER 2

Proc means returns an output of the specified summary statistics defined in the code, in this example, we specified the 5-number summary which returned the min 25 percentile, median, 75th percentile and max.

The basic output for proc univariate returns

VARIABLE	PROC MEANS	PROC UNIVARIATE	VARIABLE DISTRIBUTION
Age	Minimum – 19years Lower percentile 25% - 26.5 Median is 33.000 years Upper percentile 75% - 42.0 Maximum – 74years	0% Min – 19years 25% Q1 - 26.5 50% Median - 33.000 years 75% Q3 - 42.0 100% Max – 74years	Majority of the data is distributed to the left. The min, 25th percentile and median values are closer than the 75th and maximum values. This suggests that the shape of the distribution is skewed to the right (positive skew).
Amount of loan	Minimum – 276.00 Lower percentile 25% - 1389.50 Median - 2400 Upper percentile 75% - 4231.00 Maximum – 15945.00	0% Min – 276.00 25% Q1 – 1389.5 50% Median – 2400 75% Q3 - 4231 100% Max – 15945	Majority of the data is distributed to the left. The min, 25th percentile and median values are closer than the 75th and maximum values. This suggests that the shape of the distribution is skewed to the right (positive skew).

Duration of loan	Minimum – 4 Lower percentile 25% - 12 Upper percentile 75% - 24 Maximum – 60	Median - 18	Majority of the data is distributed to the left. The 25th percentile and 75th percentile are close to the median equally. This suggests most of the distribution is close to the centre. The extremes are spread unevenly, the maximum value is farther from the median than the min value. This suggests that the presence of outliers to the right making the distribution positively skewed.
Instalment rate as a percentage of income	Minimum – 1 Lower percentile 25% - 2 Upper percentile 75% - 4 Maximum – 4	Median - 3	The distribution is probably normal.

QUESTION 3

Univariate analysis: histograms

If a variable is normally distributed, then the average of the minimum and maximum, as well as the 25th and 75th percentile, should equal the median. Taking age as an example, these values are 46.5 and 34.25 respectively, where the median is 33, suggesting that age is skewed toward larger values.

Self-Assessment Question

The extent to which the distribution deviates from normal can be assessed visually using **proc univariate**'s **histogram** command:

```
proc univariate data=LOAN_RISK;
    histogram age;
run;
```

Use the **histogram** command to examine the distributions of *age*, *amount of loan*, *duration of loan* and *instalment rate as a percentage of income*.

Review the paper by Park (2008). Manually (*do not use SAS*) create a table showing how the values of skewness and kurtosis relate to the histograms you have drawn and the 5-number summary. (At the end of this session, you may find it an interesting challenge to create a sequence of histograms which display the 5-number summary, the skewness and the kurtosis in an inset table but this is not obligatory.)

Hint: the histogram command allows you to overlay the normal distribution:

```
proc univariate data= LOAN_RISK;
    histogram age / normal(mu= est sigma= est);
run;
```

ANSWER 3

VARIABLE	SKEWNESS	KURTOSIS
Age	From the diagram of the histogram, most of the data is to the left of the distribution. When	The peak of the distribution is a little higher than a normal distribution and has a thinner tail.

	compared to a normal distribution, this distribution suggests skewness value to be >0 which means it is positively skewed.	
Amount of Loan	From the diagram of the histogram, most of the data is to the left of the distribution. A higher percentage of the data is below the median value of 2,400. When compared to a normal distribution, this distribution is >0 which means it is positively skews with long tails.	The peak of the distribution is higher than a normal distribution. It also has a thinner tail.
Duration of Loan	From the diagram of the histogram, there isn't a pattern in the distribution as the values start increasing from 50 months. However, more of the data is distributed to the left so most of the data skewed to the left. (positive skewed).	The peak of the distribution is slightly normal when compared to a normal distribution.
Instalment	From the diagram of the histogram, most of the data is on 4.05 mark. The pattern of the distribution is uneven. The lower observations are to the left which suggest the distribution is skewed to the right (negatively skewed)	When compared to a normal distribution, the curve of the distribution is almost flat with fat tails.

QUESTION 4
Univariate analysis: theory-driven plots

Park (2008) describes the P-P plot and the Q-Q plot as theory driven visualisations of fit with the normal distribution. These can be obtained using the following code:

```
proc univariate data= LOAN_RISK;
    var age;
    ppplot age;
run;
```

or

```
proc univariate data= LOAN_RISK;
    var age;
    qqplot age / normal(mu= est sigma= est);
run;
```

Self-Assessment Question

The univariate procedure produces a table of goodness-of-fit tests for a variable by adding **normaltest** to the procedure options after **data=**. Manually (*do not use SAS*) create a table showing how the results of these tests relate to the P-P plot and the Q-Q plot for each of *age*, *amount of loan*, *duration of loan* and *instalment rate as a percentage of income*.

ANSWER 4

Variable	Shapiro-Wilk (W)	Kolmogrov-Smirnov (D)	Cramer-Von Mises (W-Sq)	Anderson-Darling (A-Sq)
Age	**Statistic value = 0.910154** The value is close to 1 which suggests a normal distribution. **p Value = <0.0001** The null hypothesis that suggest the distribution is normally distributed is rejected because the p value is less than 0.05. The distribution of the P-P plot and Q-Q plots deviate from the fitted line, which means the distribution is not normally distributed.			

QUESTION 5

Univariate analysis: extremes of the distribution

The 5 most extreme observations are listed in the usual output of proc univariate. It is sometimes useful to examine more observations than 5 and to be able to select the actual from a data set containing a primary key variable that has a unique value for each data subject.

Self-Assessment Question

Part a: write code to list the frequency of the most extreme values from the distributions of age, amount of loan, duration of loan and instalment rate as a percentage of income, using proc univariate.

Hints:

- surrounding proc univariate with ods select extreme values (before) and ods select all (after) causes the procedure to print only the table of extreme values.
- proc univariate's nextrval option controls the number of values output.

Part b: write code to print in separate tables the 10 most extreme observations from the distributions of age, amount of loan, duration of loan and instalment rate as a percentage of income.

Hints:

- surrounding proc univariate with ods select extreme jobs (before) and ods select all (after) prints only the table of extreme observations.
- proc univariate's nextrobs option controls the number of values output.

- Use of by followed by the primary key variable within proc univariate sets the key value for the list of extreme observations
- A possible strategy for this task is:

1. Create the correct output using proc univariate
2. Surround the statement with ods csvall file= '/path/to/the/CSV/file' (before) and ods csvall close (after) to output the results to a CSV file
3. Import the CSV file into a suitable SAS data set
4. Use proc SQL to create a table containing the appropriate observations
5. Use proc print to print a report of the observations

QUESTION 6
Univariate analysis: transforming data

When a measurement variable does not fit a normal distribution and you are using parametric tests (that depend on this being the case) then it is reasonable to try a transformation.

Self-Assessment Question

For each of *age*, *amount of loan*, *duration of loan* and *instalment rate as a percentage of income* that is not normally distributed, find a suitable transformation that improves the fit with a normal distribution.

Hint: Search for suitable transformations using the search terms: "statistics transforming data".