New York Air-traffic

Exploratory data analysis is compared to detective work: it is the process of gathering evidence.

- List three main questions/hypotheses you want to test about your data.
- Identify the variables that are relevant to your questions. Argue why other variables are not relevant to the questions.
- Search for evidence by visualising, transforming, and modeling your data.
- Use the evidence to answer/refine your questions, test your hypotheses, and/or generate new questions.

1. What type of variation occurs within each variable? (RDS 7.3)

(Note: you need to repeat 1.1 through 1.5 for all variables, i.e., if you have 5 variables you document the process 1.1 to 1.5 for each of those 5 variables.)

1.1 Visualising distributions (Barcharts, Histograms) (RDS 7.3.1, 7.3.2)

- Which values are the most common? Why?
- Which values are rare? Why? Does that match your expectations?
- Can you see any unusual patterns? What might explain them?
- Are there clusters in the data? If so,
  - How are the observations within each cluster similar to or different from each other?
  - How can you explain or describe the clusters?

1.2 Unusual values (RDS 7.3.3)

- Are there unusual values in the data? E.g. too large, too small, negative, etc.
- Are they outliers? How do you determine that?
- How do your distributions look like with and without outliers?

1.3 Missing values (RDS 7.4)

- Do you have missing values in your data? How do you determine that?
- How do you handle the missing values? E.g., removing, replacing with a constant value, or a value based on the distribution? Discuss your decision.
- Show how your data looks in each case
- Does converting the type of this variable help exploring the distribution of its values or identifying outliers or missing values? What type can the variable be converted to? What will the distribution look like? Please demonstrate with appropriate plots.
- What new variables do you need to create?

2. What type of covariation occurs between the variables? (RDS 7.5, 7.6)

(Note: you need to repeat 2.1 through 2.3 for all variables, i.e., if you have 5 variables you document the process for all 10 combinations of variables.)

2.1 Between a categorical and continuous variable (RDS 7.5.1)

- What type of visualization can you use?
- What patterns and relationships do you observe?
- Could the identified patterns be due to coincidence (i.e. random chance)?
- How can you describe the relationship implied by the pattern? (e.g., positive or negative correlation)
- How strong is the relationship implied by the pattern? (e.g., how strong is the correlation?)
- What other variables might affect the relationship?
- Does the relationship change if you look at individual subgroups of the data?
- Does converting the type of these variables help exploring the relationship? If so, what type? Please demonstrate with appropriate plots
- Do the observed patterns support/reject your hypotheses or answer your questions?

2.2 Between two categorical variables (RDS 7.5.2)

- What type of visualization can you use?
- What patterns and relationships do you observe?
- Could the identified patterns be due to coincidence (i.e. random chance)?
- How can you describe the relationship implied by the pattern? (e.g., positive or negative correlation)
- How strong is the relationship implied by the pattern? (e.g., how strong is the correlation?)
- What other variables might affect the relationship?
- Does the relationship change if you look at individual subgroups of the data?
- Does converting the type of these variables help exploring the relationship? If so, what type? Please demonstrate with appropriate plots
- Do the observed patterns support/reject your hypotheses or answer your questions?

2.3 Between two continuous variables (RDS 7.5.3)

- What type of visualization can you use?
- What patterns and relationships do you observe?
- Could the identified patterns be due to coincidence (i.e. random chance)?
- How can you describe the relationship implied by the pattern? (e.g., positive or negative correlation)
- How strong is the relationship implied by the pattern? (e.g., calculate the correlation)
- What other variables might affect the relationship?
- Does the relationship change if you look at individual subgroups of the data?
- Does converting the type of these variables help exploring the relationship? If so, what type? Please demonstrate with appropriate plots
- Do the observed patterns support/reject your hypotheses or answer your questions?

3. Summarize your findings about the three questions you asked at the beginning. Describe how your observations support or reject your hypotheses or answer your questions.