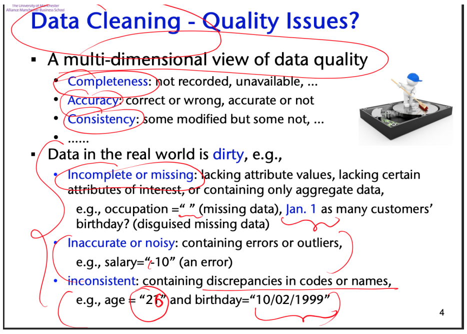
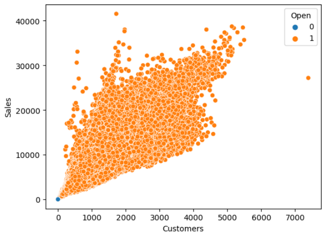
1. Data Reviewing
    a. Variables
    b. Quality
        i. A multi-dimensional view of data quality (week7)
            1. 

The University of Manchester
Alliance Manchester Business School

Data Cleaning - Quality Issues?

- A multi-dimensional view of data quality
    - Completeness: not recorded, unavailable, …
    - Accuracy: correct or wrong, accurate or not
    - Consistency: some modified but some not, …
    - ……
- Data in the real world is dirty, e.g.,
    - Incomplete or missing: lacking attribute values, lacking certain attributes of interest, or containing only aggregate data,
      e.g., occupation =“ ” (missing data), Jan. 1 as many customers’ birthday? (disguised missing data)
    - Inaccurate or noisy: containing errors or outliers,
      e.g., salary=“-10” (an error)
    - inconsistent: containing discrepancies in codes or names,
      e.g., age = “21” and birthday=“10/02/1999”

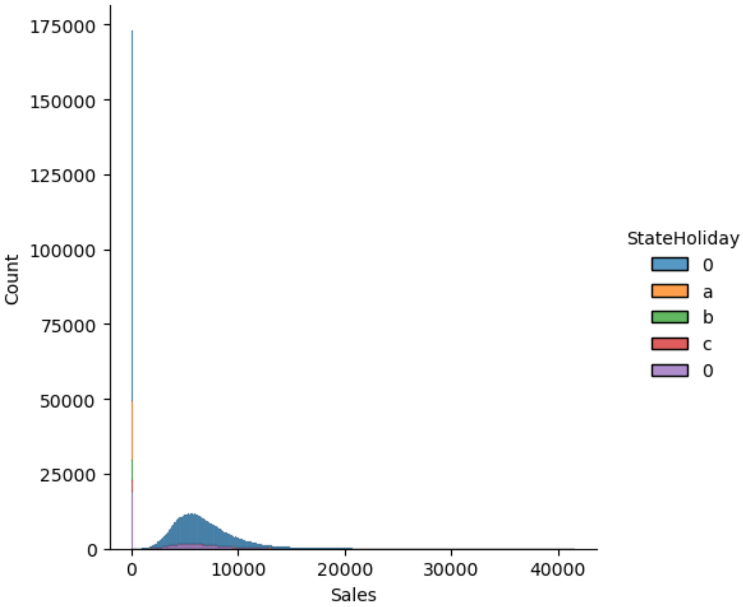
4

            2. Completeness
                a. Store data has missing values in Competition, Promo
                b. Train data has no missing values.
                c. Test data has missing values in Open
            3. Accuracy
                a. ? (Questions)
                b. Sales and Customers are always 0 if Open=0, so accurate
                    i. 

            4. Consistency
                a. Inconsistency in StateHoliday in train data.
    c. Relevance to the sales forecasting
        i. Hypothesis
            1. Time trend
                a. The sales increase with time
            2. Past sales
                a. The sales of a day are correlated with the sales of te day before.
            3. StoreType
                a. The sales of a store depend on its store type.
            4. Assortment
                a. The sales of a store depend on its assortment level.
            5. Competition
                a. The sales of a store will be lower if the store has competition.
                b. The closer the competition to the store is, the more effect it will get on its sales.

c. The effect of competition will be large immediately after its opening (Competition Shock). But the effect will decrease gradually to some extent.

6. Promo2
   a. The promotion will increase sales.
   b. The impact of promotion on sales is large in the first month of each round but gradually decreases as time goes on.
7. DayOfWeek
   a. There is a weekly cycle in sales.
   b. It depends on each store on which day it has higher sales.
8. Open
   a. A store has no sales if it is closed on that day.
9. Promo
   a. A store-specific promotion increase sales.
10. Holiday
   a. Holidays affect sales.
   b. The direction of effect (positive or negative) depends on each store.
   c.



2. Data Preprocessing
   a. Supposed model
      i. Linear regression
         1. One general model that can be applied to all stores
   b. General Preprocessing
      i. Missing Values
         1. Competition
            a. All rows with null CompetitionDistance have null CompetitionOpenMonth/Year, & min(CompetitionDistance) > 0.

               i. If CompetitionDistance == null, the store has no competition
               ii. Insert significantly large number (e.g.1e+11) to CompetitionDistance
            b. If CompetitionDistance != null & CompetitionMonth/Year == null:
               i. It means competitor opened at some point, but you cannot know when it was from the data.
               ii. In this case, HasCompetition will be always 1 in prediction part, so it might be better to train the model assuming HasCompetition = 1 over the period (?)
               iii. Assume the competition had already opened before the start of data (01/01/2013)
                  1. Insert a day <= 01/01/2013
                  2. This is because in prediction, HasCompetition is always 1.
         2. Promo2
      ii. Standardisation
   c. Making new variables
3. Questions
   a. Data Quality
      i. How can you check the accuracy of data?
   b. Data Preprocessing
      i. store_df
         1. What to do when you don’t know when the competition opened?
   c. Model
      i. How to include store_id?
         1. If one-hot encoding, too many columns→is it allowed?
         2. If average sales, it must correlated with past sales →is it allowed?
         3. If exclude store_id, 1115 models →can you combine their score in assessing?