* Be sure to clean fields we are using
* What is our use case?
* Fix what in phase 1 ?
* include phase 1 in final report
* Update and expanded version of items from phase 1
* Narrative that ties all steps together and explains motivation (use case U1), rational for design of overall workflow W and tools used
* Documenation that data quality was improved. before/ after queries
* summary of data changes resulting from overall workflow
* summary of findings, problems encountered and lessons learned including possible next steps

Supplementary materials:

Operation history: Openrefine operation history ??

queries - queries written in sql to profile the datset and check integrity constraints Queries.txt

Workflow model – Use YesWorkflow or other diagramming tools.

Raw and cleaned datasets.

Clean and validate:

availability\_by\_dte

review\_scores\_value

neighborhood

neighborhood\_cleansed

neighborhood\_group\_cleansed

property\_type

room\_type

price

U1: summarize average price and average rating by type of room per neighborhood

which neighborhood has least expensive and move expensive types of Airbnb’s by room type

Calculate which neighborhood has highest and lowest average ratings and how this corresponds to average price

Use neighbourhood\_cleansed

= we will cluster into regions and perform same analysis on region

U0 – no data clean is calculating average min night requirement for each location

dAta quality problems:

price is char

neighborhood - region

Validate the number of rows per data file

confirm no missing data points for our specific fields.

REview min values for pirce to confirm no negative price. Check range and outliers

Review min and maximum values for ratings – between 1 and 10

group property type and confirm number of distinct property types

Validate neighbourhood values in file exist in neighborhoods data set

Check for number of reviews in the past 30 days < total reviews

link calendar dataset to review avg price over available dates and confirm that it is in the range of price in listing data set

regions ?

neighbrourhood -> region city

city, region

**REATE TABLE listings (id, listing\_url, scrape\_id, last\_scraped, name, description, neighborhood\_overview, picture\_url, host\_id, host\_url, host\_name, host\_since, host\_location, host\_about, host\_response\_time, host\_response\_rate, host\_acceptance\_rate, host\_is\_superhost, host\_thumbnail\_url, host\_picture\_url, host\_neighbourhood, host\_listings\_count, host\_total\_listings\_count, host\_verifications, host\_has\_profile\_pic, host\_identity\_verified, neighbourhood, neighbourhood\_cleansed, neighbourhood\_group\_cleansed, latitude, longitude, property\_type, room\_type, accommodates, bathrooms, bathrooms\_text, bedrooms, beds, amenities, price, minimum\_nights, maximum\_nights, minimum\_minimum\_nights, maximum\_minimum\_nights, minimum\_maximum\_nights, maximum\_maximum\_nights, minimum\_nights\_avg\_ntm, maximum\_nights\_avg\_ntm, calendar\_updated, has\_availability, availability\_30, availability\_60, availability\_90, availability\_365, calendar\_last\_scraped, number\_of\_reviews, number\_of\_reviews\_ltm, number\_of\_reviews\_l30d, first\_review, last\_review, review\_scores\_rating, review\_scores\_accuracy, review\_scores\_cleanliness, review\_scores\_checkin, review\_scores\_communication, review\_scores\_location, review\_scores\_value, license, instant\_bookable, calculated\_host\_listings\_count, calculated\_host\_listings\_count\_entire\_homes, calculated\_host\_listings\_count\_private\_rooms, calculated\_host\_listings\_count\_shared\_rooms, reviews\_per\_month)''')**