Appendix 1: Data preparation change log

File	Variable	Comment	Action
listings, calendar, reviews	id, host id, scrape id	These variables are unique identifier that have no statistics meaning.	Change data type from "Numeric" to "Character" and choose "Nominal" as modeling type.
listings	host_response_rate,	Rate should be in percentage form to get statistics.	Change its data type from "Character" to "Numeric" and choose "Continuous" as modeling type.
listings, reviews	summary, space, description, neighborhood_overview, notes, transit, access, interaction, house_rules, comments	These are unstructured text.	change modeling type to "Unstructured Text" by standardize their attributes.
listings	Latitude, longitude	These variables refer to location.	Correct data format as "latitude DMS" and "Longitude DMS".
	price, weekly_price, monthly_price, security_deposit, cleaning_fee, extra_people	Price variables should have specific format.	Standardize format as "Singapore Dollar ($)" and fix at 0 decimal places.
	Thumbnail_url, medium_url, xl_picture_url, state, square_feet, license, jurisdiction_names	Overmuch missing values (cover 90% or more)	Excluded and hid these columns.
	Host_name, host_since, host_response_time, host_response_rate, hose_acceptance_rate, host_is_superhost, host_thumbnail_url, host_picture_url, etc.	The same 82 observations missed of these variables related to host.	Delete these 82 rows.
	host_listings_count, host_total_listings_count	Have unreasonable o values.	Insert a column to calculate the correct number of listings each host own and replace "0" with the correct value.

	host_since	The dates in “host_since” should be earlier than those in “first_review”.	Replace this abnormal data with the date of the first review received by the host from all the listings owned.
listings, calendar	price, adjust_price	Price should be greater than 0.	Delete these 3 rows valued 0.
	Price, adjust_price	Overmuch outliers would decrease accuracy of analysis.	Delete value greater than 90% quantile $290.
	host_is_superhost, host_has_profile_pic, host_identity_verified, etc.	Boolean value can be modified to number 0 and 1.	Change Boolean value to number “0” and “1”. Standardize their data type as “Numeric” and modeling type as “Continuous”.
listings	host_location, cancellation_policy	Too many different categories.	Group similar values.
	reviews_per_month	Different decimal places.	Fix data format at 2 decimals.
	scrape_id, last_scraped, experiences_offered, picture_url, host_acceptance_rate, host_thumbnail_url, host_picture_url, street, city, market, smart_location, country_code, country, etc.	Useless variables	Exclude and hide these columns.