## Data Science – Data Prep with R – Quick Reference

PROFILE DATASET	
Volume	df2 <- df %>% summarize (count = n())
Velocity	df2 <- df %>% group_by (date1) %>% summarize (count = n())
Attribute Selection	df2 <- df %>% select (c ('attr1' , 'attr2'))
Incomplete Records	df2 <- df %>% filter (!is.na (attr1))

VALIDATE ATTRIBUTES	
Domain	distinct(df,attr1)
Missing Values	df2 <- df %>% filter(is.na (attr1))
Range	summary (df)
Data Types	str (df)
Outliers	summary (df); hist (df\$attr1)
Distribution	hist (df\$attr1)

STAN	Data Types mutate (attr1 = as.integer (attr1),	
Data Types	mutate (attr1 = as.integer (attr1), attr2 = factor (attr2), date1 = as.Date (date1))	
Patterns	mutate (attr1 = if_else (attr1 == 'Street', 'St', attr1)	
Formatting	mutate (attr1 = toupper (attr1))	
Scaling	mutate (attr1=scale (attr1))	

CLEAN ATTRIBUTES	
Outliers	mutate (attr1 = if_else (attr1 > 1000
(Quantitative)	attr1 < 0, NA, attr1)
Missing Values	mutate(attr1 = if_else (is.na(attr1),
(At Random)	mean(attr1, na.rm=TRUE), attr1))
Missing Values	mutate(attr1 = if_else
(Not at Random)	(is.na (attr1), 1, attr1)
Incorrect	mutate(attr1 = if_else
Values	(attr1 == 'bad', 'good', attr1))

DERIVE ATTRIBUTES	
Buckets/Binning	mutate (attr1_bin = cut(x = attr1, breaks = c(0,50,100)))
Date Parts	mutate (month = format(date1, format = "%m")
Date Difference	mutate (elapsed_days = difftime (date1, date2, units = 'days')
Last Period	mutate (last_year = as.numeric (format(date1, "%Y"))-1
Dummy Encoding (One Hot)	mutate (gender_male = if_else (attr1 == 'male', 1, 0)

COMBINE DATASETS	
Join Horizontally (Full Match)	df3 <- inner_join (x=df1, y=df2, by='attr1')
Join Horizontally (Optional Match)	df3 <- left_join (x=df1, y=df2, by='attr1')
Union Vertically (Deduplicate)	df3 <- rbind (df1, df2) df4 <- df3 [match (unique (df3\$attr1), df3\$attr1), ]
Union Vertically (No Deduplicate)	df3 <- rbind (df1, df2)

SPLIT DATASETS	
Simple Filter	df2 <- df %>% filter(attr1>5)
Filter Based on	df2 <- df %>%
Aggregation	filter(attr1 > mean(attr1))
Sampling	set.seed (100)
(Random)	df2 <- sample_n (df, 1000)
Sampling	df2 <- df %>%
(Non-Random)	filter (ntile (attr1, 4) == 4)

CREATE INTERFACE		
Python	library (reticulate)	
SQL	library (dbi)	

All items assume **dplyr** is loaded from tidyverse package. **df** is a dataframe with attributes attr1, attr2, date1, date2.