Assignment 4, Part 2

Brianna Kincaid

1 Table 4 to Table 6

1.1 Load Data

```
> library(foreign)
> library(stringr)
> library(plyr)
> library(reshape2)
> source("xtable.r")
> pew <- read.spss("pew.sav")
> pew <- as.data.frame(pew)
```

religion	<$10k	$10-20k	$20-30k	$30-40k	$40-50k	$50-75k
Agnostic	27	34	60	81	76	137
Atheist	12	27	37	52	35	70
Buddhist	27	21	30	34	33	58
Catholic	418	617	732	670	638	1116
s Don?t know/refused	15	14	15	11	10	35
Evangelical Prot	575	869	1064	982	881	1486
Hindu	1	9	7	9	11	34
Historically Black Prot	228	244	236	238	197	223
Jehovah's Witness	20	27	24	24	21	30
Jewish	19	19	25	25	30	95

Table 4: The first ten rows of data on income and religion from the Pew Forum. Three columns, $\$75 - 100k$ and $> 150k$, have been omitted.

1.2 Tidy Data

```
> religion <- pew[c("q16", "reltrad", "income")]
> religion$reltrad <- as.character(religion$reltrad)
> religion$reltrad <- str_replace(religion$reltrad, " Churches", "")
> religion$reltrad <- str_replace(religion$reltrad, " Protestant", " Prot")
> religion$reltrad[religion$q16 == " Atheist (do not believe in God) "] <- "Atheist"
> religion$reltrad[religion$q16 == " Agnostic (not sure if there is a God) "] <- "Agnostic"
> religion$reltrad <- str_trim(religion$reltrad)
> religion$reltrad <- str_replace_all(religion$reltrad, " \\(.*?\\)", "")
> religion$income <- c("Less than $10,000" = "<$10k",
+                      "10 to under $20,000" = "$10-20k",
+                      "20 to under $30,000" = "$20-30k",
+                      "30 to under $40,000" = "$30-40k",
+                      "40 to under $50,000" = "$40-50k",
+                      "50 to under $75,000" = "$50-75k",
+                      "75 to under $100,000" = "$75-100k",
+                      "100 to under $150,000" = "$100-150k",
+                      "$150,000 or more" = ">150k",
+                      "Don't know/Refused (VOL)" = "Don't know/refused")[religion$income]
```

```
> religion$income <- factor(religion$income, levels = c("<$10k", "$10-20k", "$20-30k", "$30-40k", "$40-50k
+                                                        "$75-100k", "$100-150k", ">150k", "Don't know/refus
> counts <- count(religion, c("reltrad", "income"))
> names(counts)[1] <- "religion"
> xtable(counts[1:10, ], file = "pew-clean.tex")
```

religion	income	freq
Agnostic	<$10k	27
Agnostic	$10-20k	34
Agnostic	$20-30k	60
Agnostic	$30-40k	81
Agnostic	$40-50k	76
Agnostic	$50-75k	137
Agnostic	$75-100k	122
Agnostic	$100-150k	109
Agnostic	>150k	84
Agnostic	Don't know/refused	96

Table 6: The first ten rows of the tidied Pew survery dataset on income and religion. The column has been renamed to income, and value to freq.

2 Table 7 to Table 8

2.1 Load Data

```
> raw <- read.csv("billboard.csv")
```

year	artist	track	time	date.entered	wk1	wk2	wk3
2000	2 Pac	Baby Don't Cry	4:22	2000-02-26	87	82	72
2000	2Ge+her	The Hardest Part Of ...	3:15	2000-09-02	91	87	92
2000	3 Doors Down	Kryptonite	3:53	2000-04-08	81	70	68
2000	98^0	Give Me Just One Nig...	3:24	2000-08-19	51	39	34
2000	A*Teens	Dancing Queen	3:44	2000-07-08	97	97	96
2000	Aaliyah	I Don't Wanna	4:15	2000-01-29	84	62	51
2000	Aaliyah	Try Again	4:03	2000-03-18	59	53	38
2000	Adams, Yolanda	Open My Heart	5:30	2000-08-26	76	76	74

Table 7: The first eight Billboard top hits for 2000. Other columns not shown are wk4, wk5, ..., wk75.

2.2 Tidy Data

```
> options(stringsAsFactors = FALSE)
> library(lubridate)
> library(reshape2)
> library(stringr)
> library(plyr)
> source("xtable.r")
> raw <- raw[, c("year", "artist.inverted", "track", "time", "date.entered", "x1st.week", "x2nd.week", "x3
> names(raw)[2] <- "artist"
> raw$artist <- iconv(raw$artist, "MAC", "ASCII//translit")
> raw$track <- str_replace(raw$track, " \\(.*?\\)", "")
> names(raw)[-(1:5)] <- str_c("wk", 1:76)
> raw <- arrange(raw, year, artist, track)
> long_name <- nchar(raw$track) > 20
> raw$track[long_name] <- paste0(substr(raw$track[long_name], 0, 20), "...")
> xtable(raw[c(1:3, 6:10), 1:8], "billboard-raw.tex")
> clean <- melt(raw, id = 1:5, na.rm = T)
> clean$week <- as.integer(str_replace_all(clean$variable, "[^0-9]+", ""))
> clean$variable <- NULL
> clean$date.entered <- ymd(clean$date.entered)
> clean$date <- clean$date.entered + weeks(clean$week - 1)
> clean$date.entered <- NULL
> clean <- rename(clean, c("value" = "rank"))
> clean <- arrange(clean, year, artist, track, time, week)
> clean <- clean[c("year", "artist", "time", "track", "date", "week", "rank")]
> clean_out <- mutate(clean,
+                     date = as.character(date))
> xtable(clean_out[1:15, ], "billboard-clean.tex")
```

2.2.1 Normalization

```
> song <- unrowname(unique(clean[c("artist", "track", "time")]))
> song$id <- 1:nrow(song)
> narrow <- song[1:15, c("id","artist", "track", "time")]
> xtable(narrow, "billboard-song.tex")
> rank <- join(clean, song, match = "first")
> rank <- rank[c("id", "date", "rank")]
> rank$date <- as.character(rank$date)
> xtable(rank[1:15, ], "billboard-rank.tex")
```

year	artist	time	track	date	week	rank
2000	2 Pac	4:22	Baby Don't Cry	2000-02-26	1	87
2000	2 Pac	4:22	Baby Don't Cry	2000-03-04	2	82
2000	2 Pac	4:22	Baby Don't Cry	2000-03-11	3	72
2000	2 Pac	4:22	Baby Don't Cry	2000-03-18	4	77
2000	2 Pac	4:22	Baby Don't Cry	2000-03-25	5	87
2000	2 Pac	4:22	Baby Don't Cry	2000-04-01	6	94
2000	2 Pac	4:22	Baby Don't Cry	2000-04-08	7	99
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-02	1	91
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-09	2	87
2000	2Ge+her	3:15	The Hardest Part Of ...	2000-09-16	3	92
2000	3 Doors Down	3:53	Kryptonite	2000-04-08	1	81
2000	3 Doors Down	3:53	Kryptonite	2000-04-15	2	70
2000	3 Doors Down	3:53	Kryptonite	2000-04-22	3	68
2000	3 Doors Down	3:53	Kryptonite	2000-04-29	4	67
2000	3 Doors Down	3:53	Kryptonite	2000-05-06	5	66

Table 8: First fifteen rows of the tidied Billboard dataset. The date column does not appear in the original table, but can be computed from date.entered and week.