EDA

```
survivor_data_final = read.csv("data/survivor_data_final.csv")
```

Summary

```
survivor_data_final %>%
  select(gender, poc, personality_type_binary, age_during_show, days_survived, region) %>%
  tbl_summary(type = list(gender ~ "categorical",
                          poc ~ "categorical",
                          personality_type_binary ~ "categorical",
                          region ~ "categorical",
                          age_during_show ~ "continuous",
                          days_survived ~ "continuous"),
              statistic = list(all_continuous() ~ "{mean} ({sd})"),
              digits = all_continuous() ~ 1,
              label = list(c(gender) ~ "Gender",
                           c(poc) ~ "Race Identifier",
                           c(personality_type_binary) ~ "Personality Type",
                           c(region) ~ "Region",
                           c(age_during_show) ~ "Age During Show (Years)",
                           c(days_survived) ~ "Survival Time on Show (Days)")) %>%
  bold_labels()
```

Characteristic	**N = 728**
Gender	
Female	356 (49%)
Male	368 (51%)
Unknown	4
Race Identifier	
POC	199 (27%)
White	525 (73%)
Unknown	4
Personality Type	
Extrovert	401 (56%)
Introvert	320 (44%)
Unknown	7
Age During Show (Years)	33.4 (10.1)
Survival Time on Show (Days)	23.9 (12.1)
Region	
Midwest	99 (14%)
Northeast	153 (21%)
South	207 (28%)
West	269 (37%)

Note: N = 728 refers to the total count of records (i.e. contestant occurrences) in `survivor_data_final`; distinct persons may be listed in multiple records, across seasons and/or within seasons.

```
## Personality type
survivor_data_final %>%
  group_by(personality_type_binary) %>%
  summarize(n_personality_dist = n_distinct(full_name),
            n_personality_occ = n(),
            mean_days_survived = mean(days_survived)) %>%
    na.omit()  %>%
  knitr::kable(digits = 1, col.names = c("Personality Type", "Distinct Persons", "Contestant Occurrences
```

Personality Type	Distinct Persons	Contestant Occurrences	Mean Days Survived
Extrovert	309	401	24.0
Introvert	271	320	23.6

```
## POC Status
survivor_data_final %>%
  group_by(poc) %>%
  summarize(n_poc_dist = n_distinct(full_name),
            n_poc_occ = n(),
            mean_days_survived = mean(days_survived, na.rm = TRUE)) %>%
    na.omit()  %>%
  knitr::kable(digits = 1, col.names = c("POC Status", "Distinct Persons", "Contestant Occurrences", "M
```

POC Status	Distinct Persons	Contestant Occurrences	Mean Days Survived
POC	164	199	22.6
White	418	525	24.3

```
## Gender
survivor_data_final %>%
  group_by(gender) %>%
  summarize(n_gender_dist = n_distinct(full_name),
            n_gender_occ = n(),
            mean_days_survived = mean(days_survived, na.rm = TRUE)) %>%
    na.omit()  %>%
  knitr::kable(digits = 1, col.names = c("Gender", "Distinct Persons", "Contestant Occurrences", "Mean
```

Gender	Distinct Persons	Contestant Occurrences	Mean Days Survived
Female	292	356	23.1
Male	290	368	24.5

```
## Region
survivor_data_final %>%
  group_by(region) %>%
  summarize(n_gender_dist = n_distinct(full_name),
            n_gender_occ = n(),
```

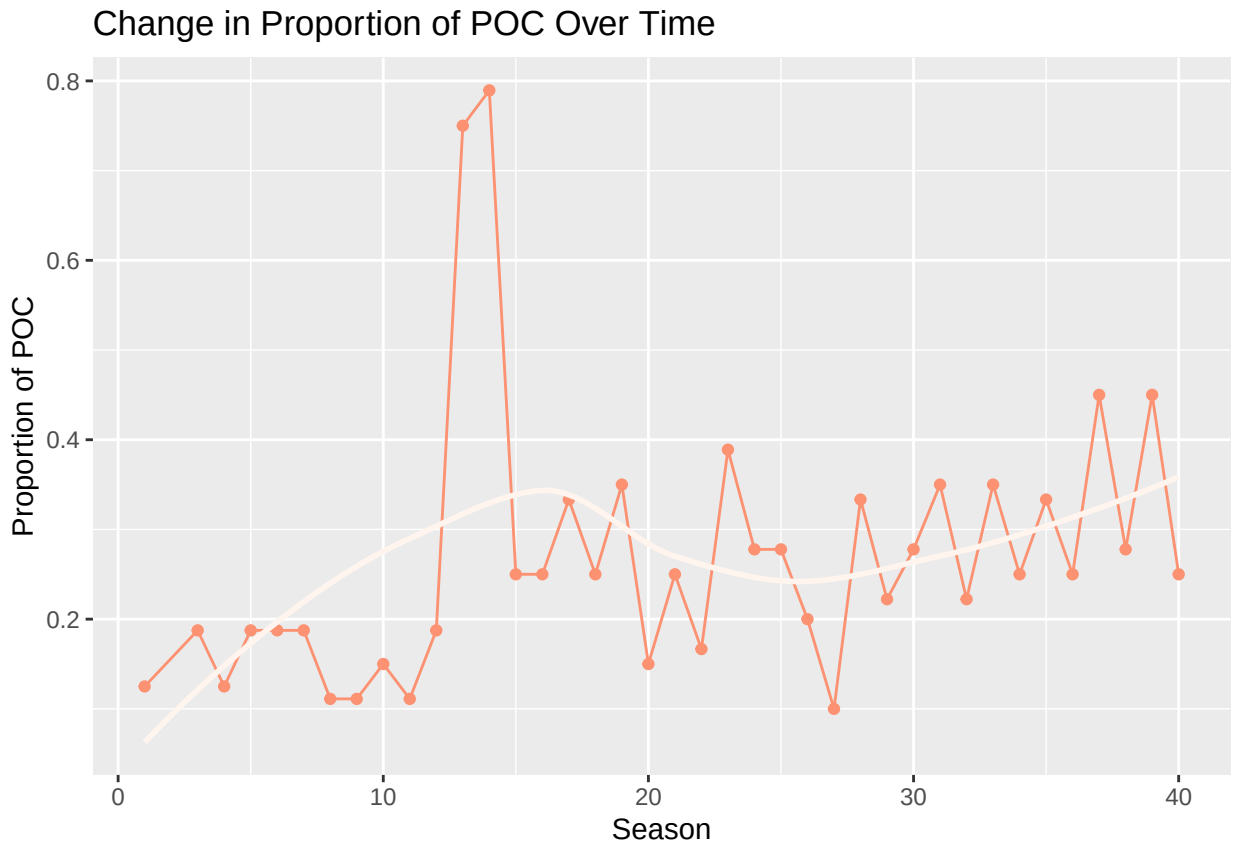
```
            mean_days_survived = mean(days_survived, na.rm = TRUE)) %>%
    na.omit()  %>%
  knitr::kable(digits = 1, col.names = c("Region", "Distinct Persons", "Contestant Occurrences", "Mean
```

Region	Distinct Persons	Contestant Occurrences	Mean Days Survived
Midwest	84	99	24.4
Northeast	122	153	25.0
South	178	207	22.7
West	218	269	23.8

Note: We report both distinct person counts and contestant occurrences by personality type, POC status, and gender.

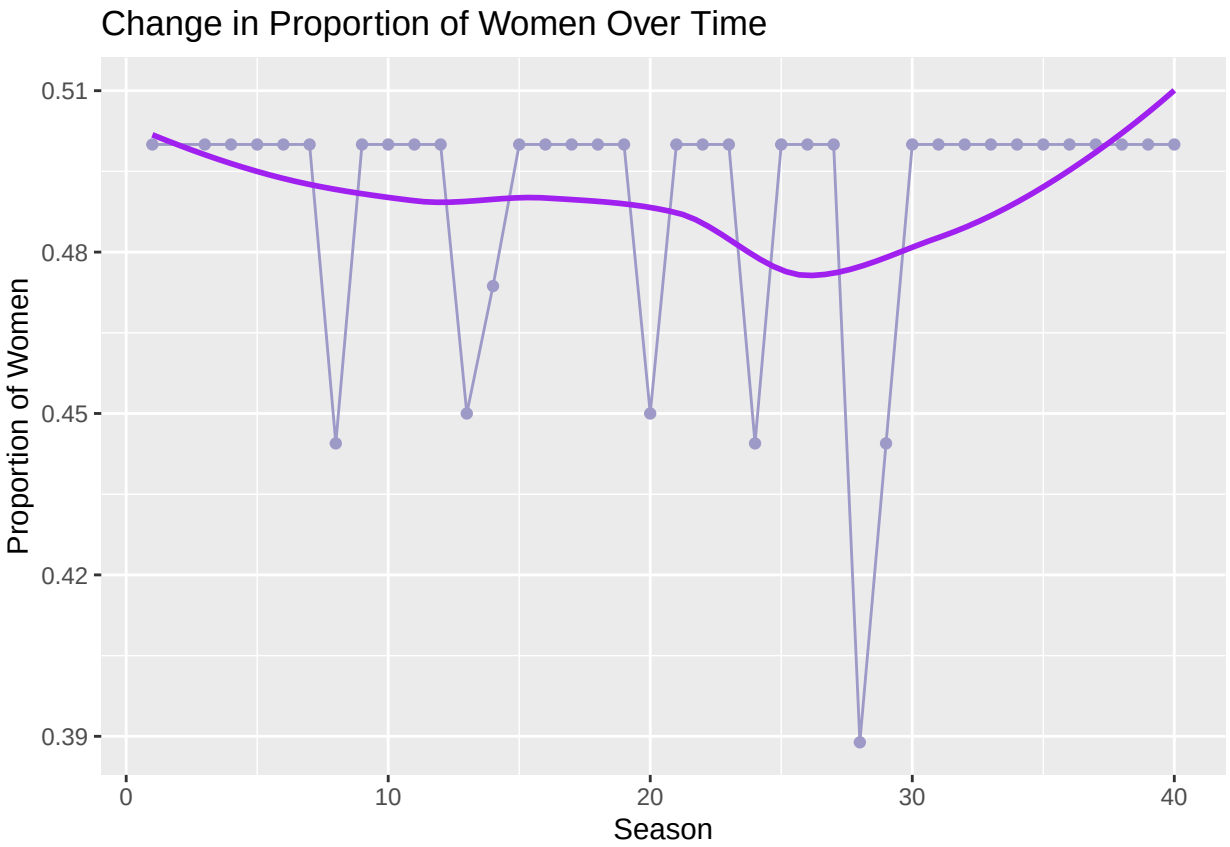
POC and Gender Representation Across Seasons

```
fill_color = brewer.pal(9,"Reds")[4]
survivor_poc_over_time = survivor_data_final %>%
  group_by(version_season, poc) %>%
  summarize(count = n_distinct(full_name)) %>%
  mutate(freq = count / sum(count)) %>%
  filter(poc == "POC") %>%
  separate(col = version_season, into = c('NA', 'season'), sep = 2) %>%
  dplyr::select(-"NA") %>%
  mutate(season = as.numeric(season))

ggplot(data = survivor_poc_over_time, aes(x = season, y = freq, group = 1)) +
  geom_line(color = fill_color) +
  geom_point(color = fill_color) +
  geom_smooth(se = FALSE, color = "seashell") +
  ggtitle("Change in Proportion of POC Over Time ") +
  xlab("Season") + ylab("Proportion of POC")
```



Note: Distinct person counts by POC status.

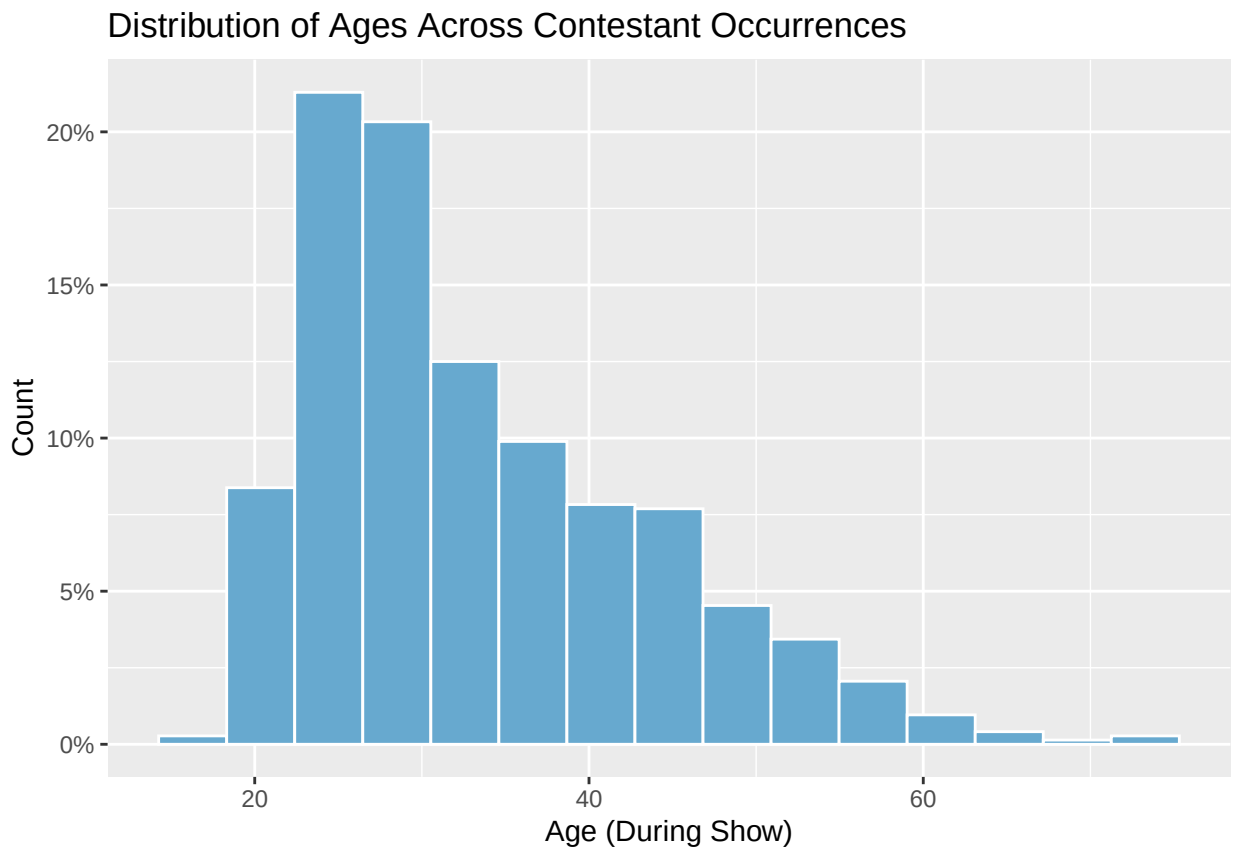
```
fill_color = brewer.pal(9,"Purples")[5]
survivor_gender_over_time = survivor_data_final %>%
  group_by(version_season, gender) %>%
  summarize(count = n_distinct(full_name)) %>%
  mutate(freq = count / sum(count)) %>%
  filter(gender == "Female") %>%
  separate(col = version_season, into = c('NA', 'season'), sep = 2) %>%
  dplyr::select(-"NA") %>%
  mutate(season = as.numeric(season))

ggplot(data = survivor_gender_over_time, aes(x = season, y = freq, group = 1)) +
  geom_line(color = fill_color) +
  geom_point(color = fill_color) +
  geom_smooth(se = FALSE, color = "purple") +
  ggtitle("Change in Proportion of Women Over Time") +
  xlab("Season") + ylab("Proportion of Women")
```



Note: Distinct person counts by gender.

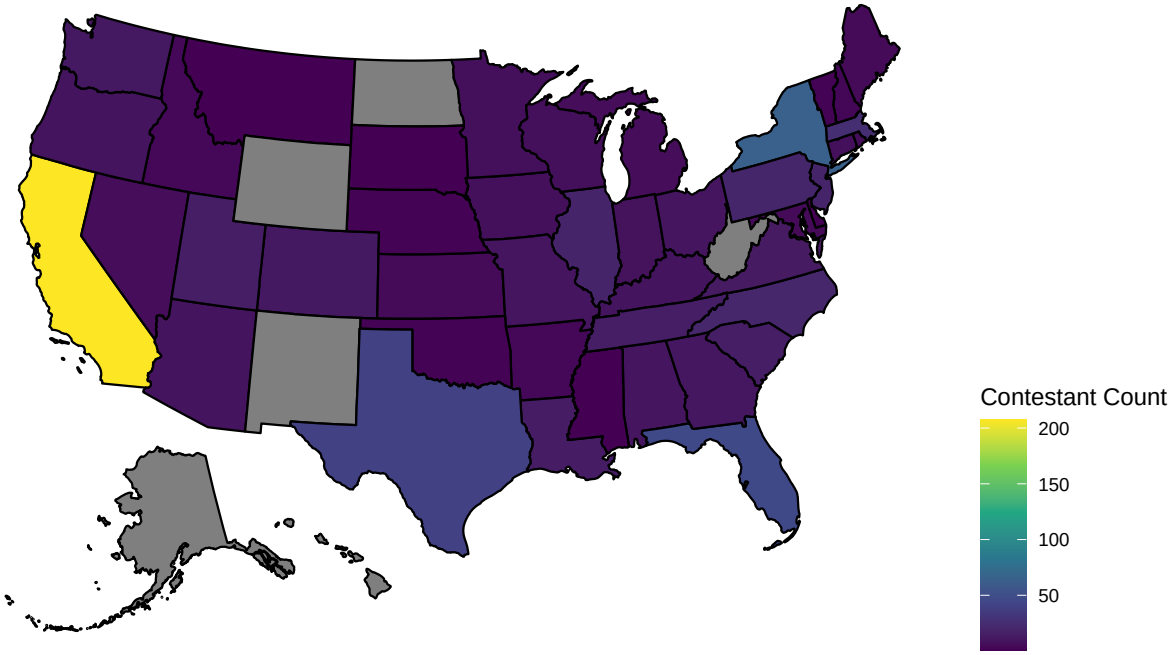
Concentration of Contestants by Age and Geography

```
fill_color = brewer.pal(9,"PuBuGn")[5]
ggplot(survivor_data_final, aes(x = age_during_show)) +
  geom_histogram(aes(y = after_stat(count/sum(count))), bins = 15, fill = fill_color, col = "white") +
  scale_y_continuous(labels = scales::percent) +
  ggtitle("Distribution of Ages Across Contestant Occurrences") +
  xlab("Age (During Show)") + ylab("Count")
```



Note: Since contestants can re-appear across seasons at different ages, we rely on discrete records from `survivor_data_final` (i.e. contestant occurrences) as the unit of analysis for this plot in order to ensure comprehensiveness of age data.

```
survivor_state = survivor_data_final %>%
  group_by(state) %>%
  summarize(n = n())

plot_usmap(
  data = survivor_state, values = "n", lines = "blue"
) +
  scale_fill_continuous(type = "viridis", name = "Contestant Count", label = scales::comma
  ) +
  labs(title = "US States", subtitle = "Geographic Distribution of Contestants") +
  theme(legend.position = "right")
```



Notes: (i) Seasons 2, 41, 42, and 43 have been removed from the exploratory analysis due to inconsistent number of days. (ii) Since contestants can re-appear across seasons with different states of residence, we similarly rely on discrete records from `survivor_data_final` (i.e. contestant occurrences) as the unit of analysis for this plot in order to ensure comprehensiveness of location data.