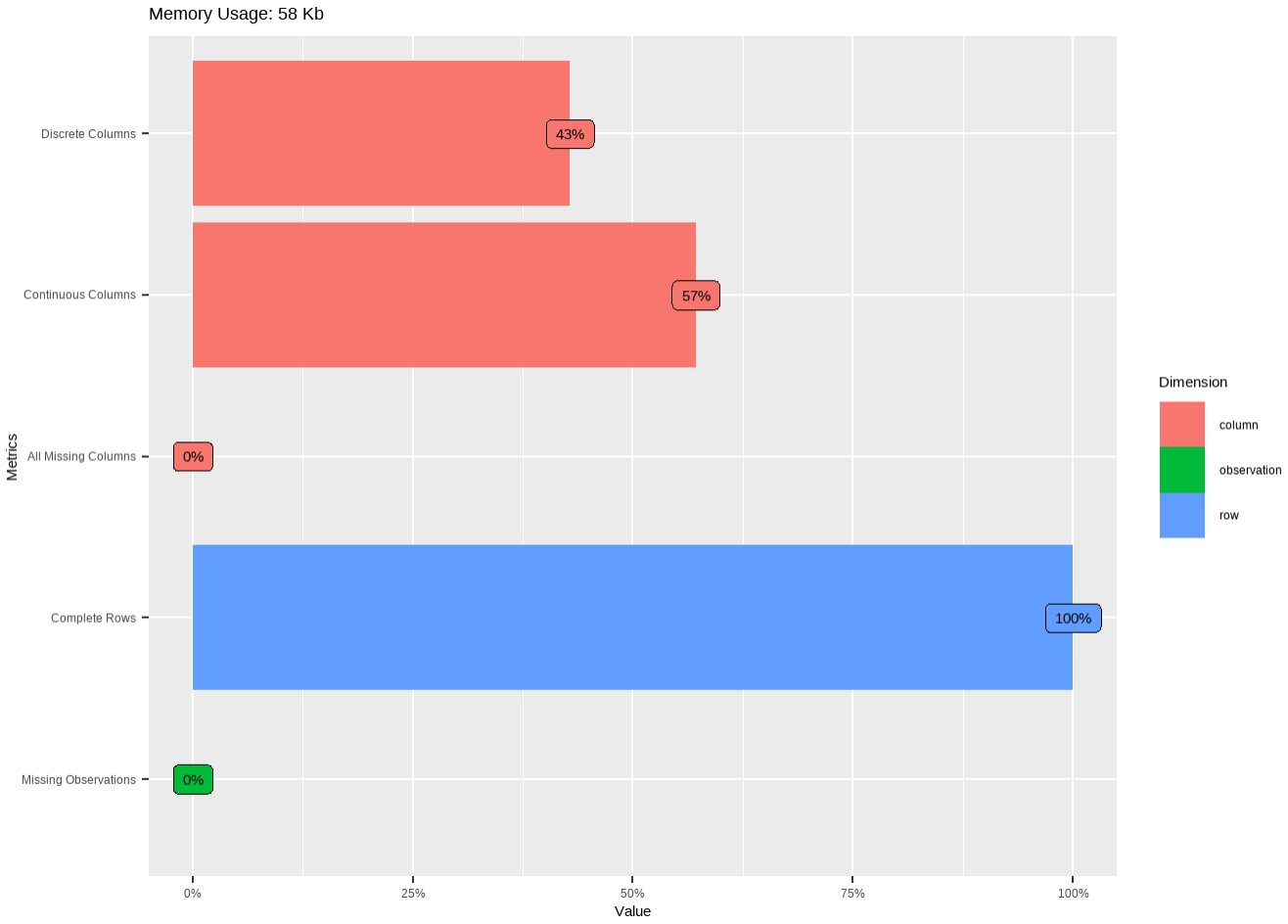
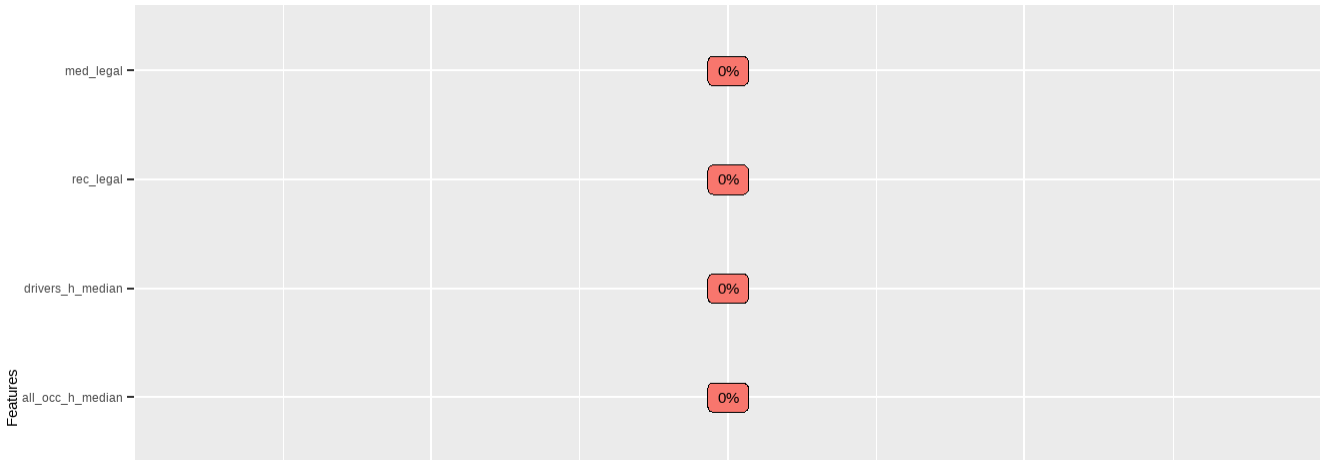
Marijuana Evaluation

We can first check some basics about the data set we imported. Do we have any missing values? How many observations of what kinds of variables do we have? What are the distributions of our variables?

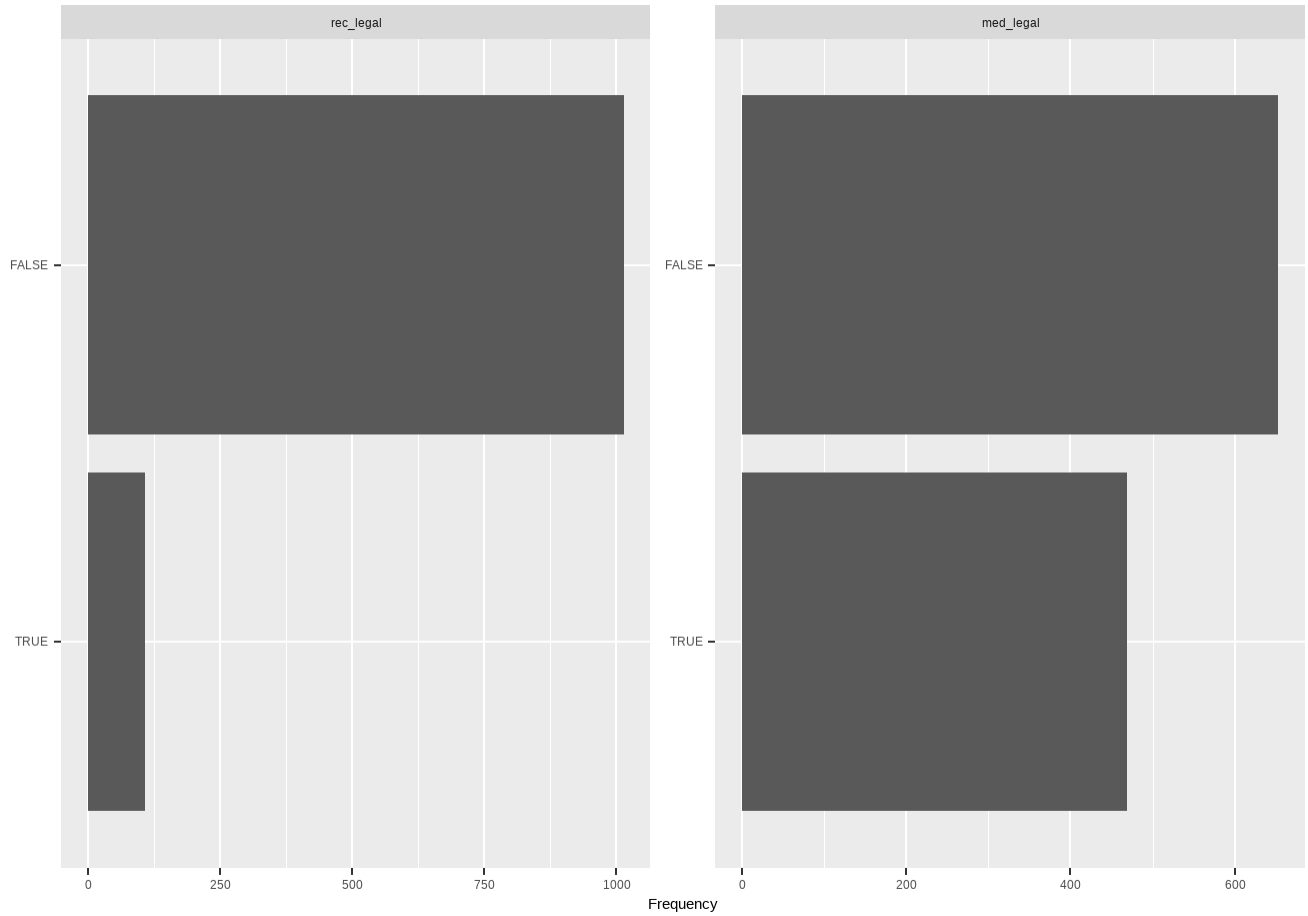
```
plot_intro(data)
```



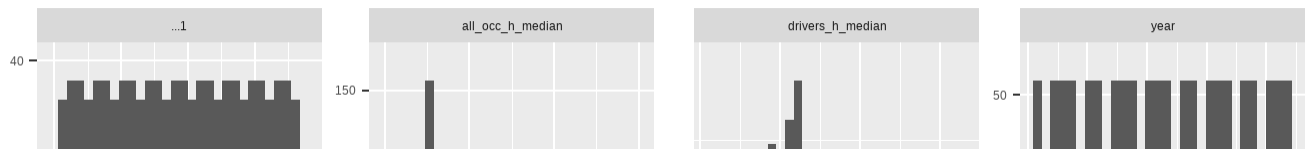

```
plot_missing(data)
```

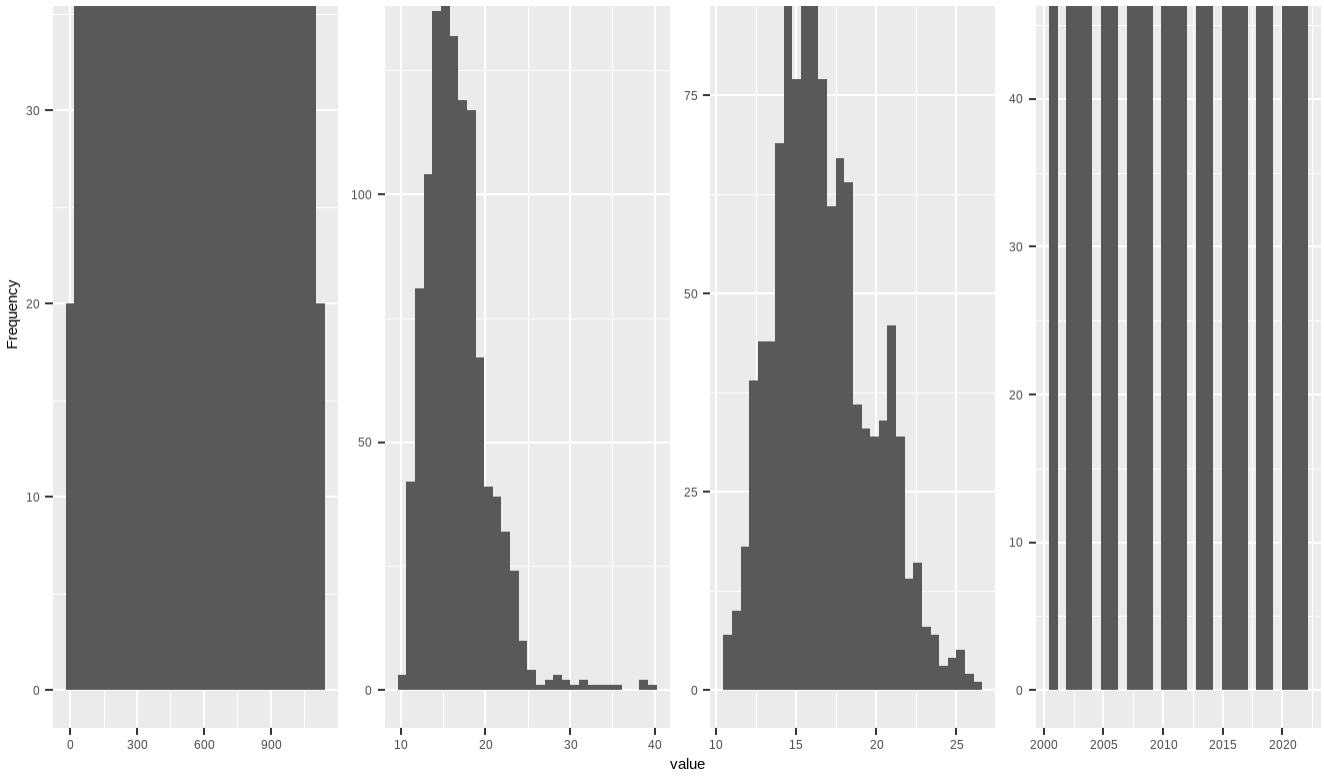



```
plot_bar(data)
```




```
plot_histogram(data)
```

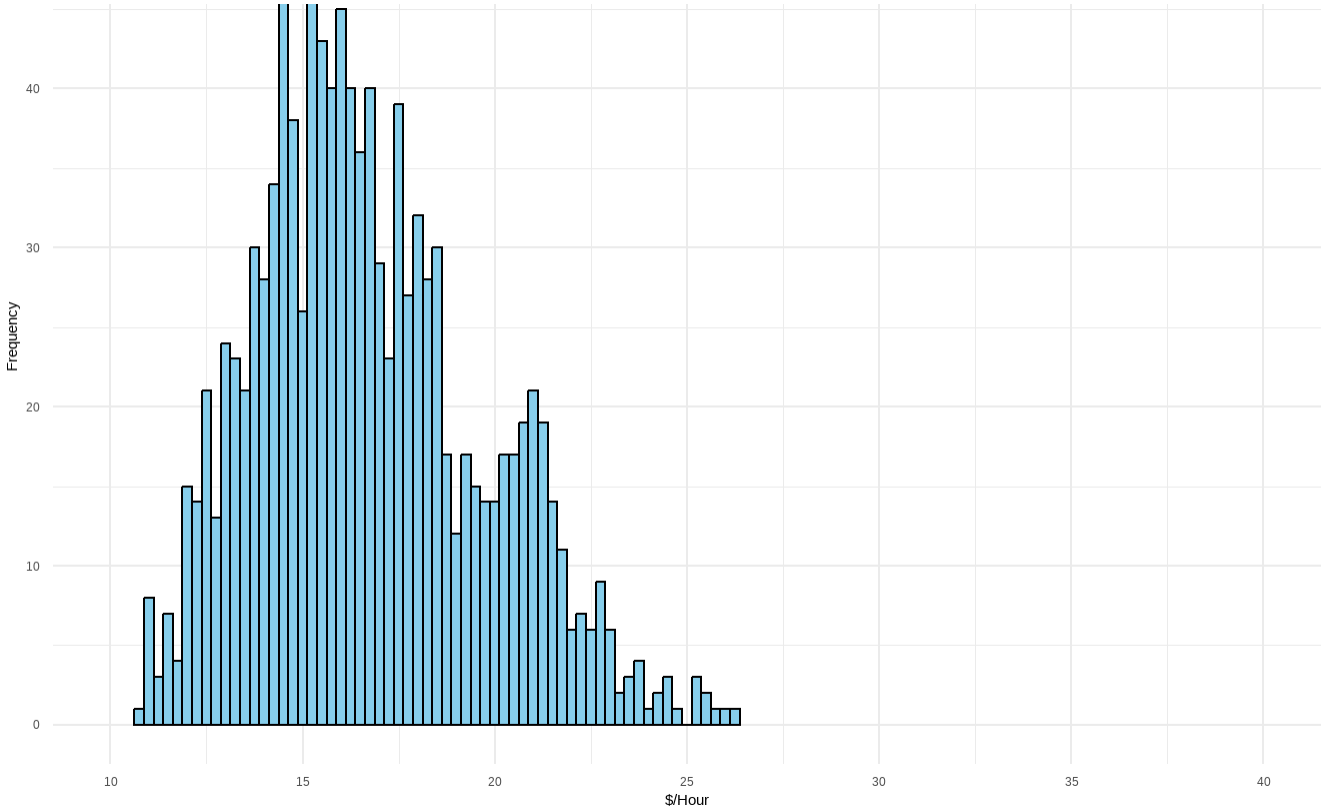




Our quick and dirty EDA has shown that we have no missing data. We see we have heavily imbalanced classes, with medicinal marijuana being illegal in a slight majority of state-years and recreational marijuana being illegal in a significant majority of state-years. We will have to keep this in mind with the final statistical analysis we choose to run and ensure that it is compatible with these unbalanced classes.

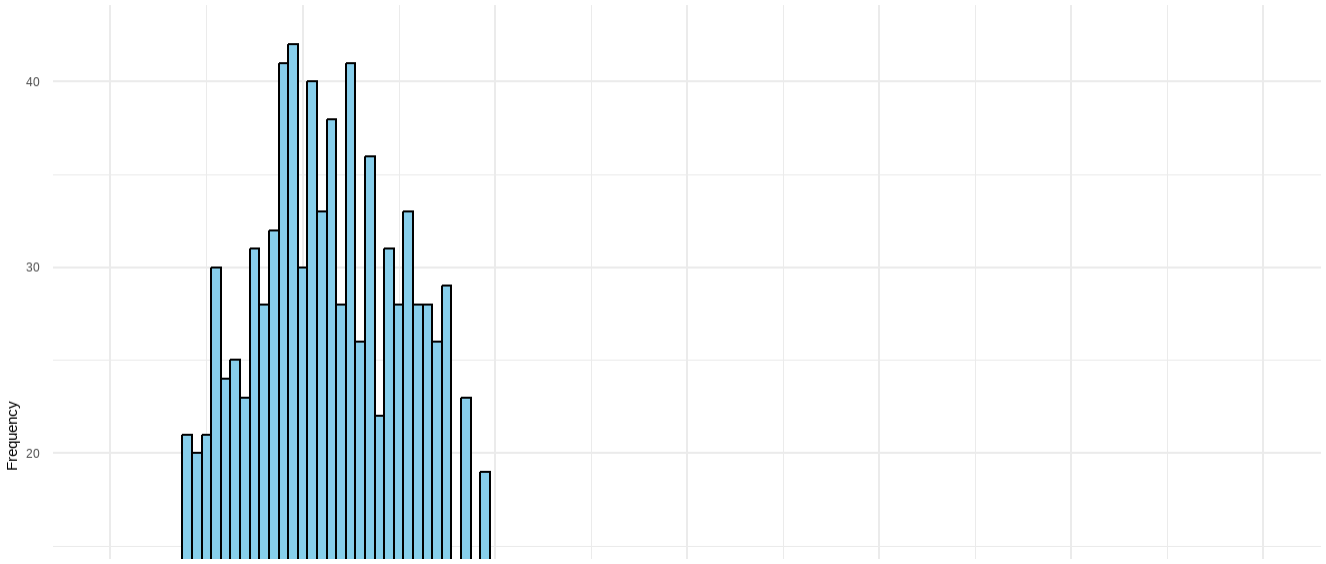
Our package for easy EDA, DataExplorer, has turned up histograms for the index and year columns which are not useful. However, we do see histograms for the median wage of both drivers and all occupations. These are scaled with different x-axes, so we'll remake these plots manually in ggplot to get a better idea of what's going on. It's worth observing that both appear to be right-skewed, which is not unexpected with wage data which has an absolute floor but no absolute ceiling. The "all occupations" graph seems to have a bigger tail; we can quantify this later.

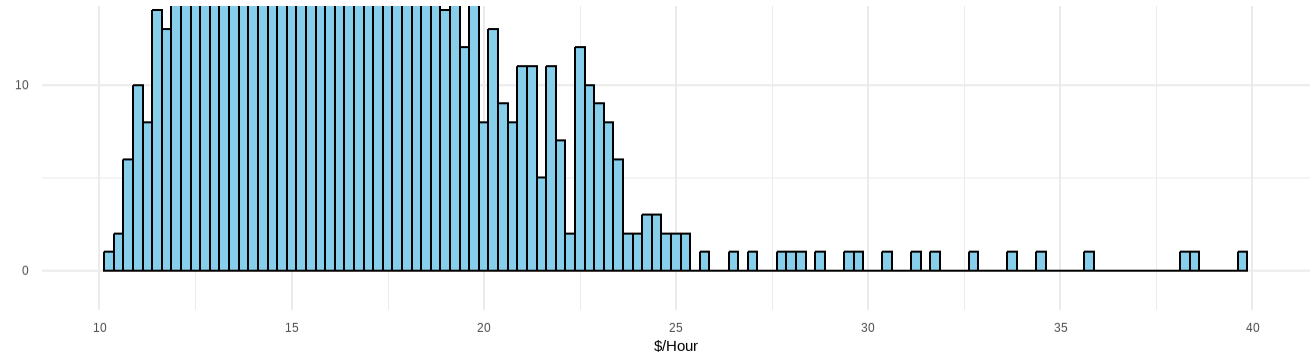
```
ggplot(data, aes(x = drivers_h_median)) +
  geom_histogram(binwidth = .25, fill = "skyblue", color = "black") +
  scale_x_continuous(breaks = seq(0, 40, by = 5)) +
  labs(
    title = "Histogram of driver state median wages 2001-20222",
    x = "$/Hour",
    y = "Frequency"
  ) +
  coord_cartesian(xlim = c(10, 40)) +
  theme_minimal()
```




```
ggplot(data, aes(x = all_occ_h_median)) +
  geom_histogram(binwidth = .25, fill = "skyblue", color = "black") +
  scale_x_continuous(breaks = seq(0, 40, by = 5)) +
  labs(
    title = "Histogram of state median wages across all occuptations 2001-20222",
    x = "$/Hour",
    y = "Frequency"
  ) +
  coord_cartesian(xlim = c(10, 40)) +
  theme_minimal()
```

Histogram of state median wages across all occuptations 2001-20222





We can see now that the distributions of wages look very similar across drivers and all occupations, with the exception of the longer tail on all occupations suggesting there's a very small number of state-years that have very high median wages.

Since we know that we have no missing data, we'll begin by looking at some summary statistics. We'll then decide what statistical test is appropriate.

```
summary_table <- data |>
  group_by(med_legal, rec_legal) |>
  summarize(
    mean_all = mean(all_occ_h_median), n = n(),
    mean_drivers = mean(drivers_h_median), n = n())

summary_table
```

```
# A tibble: 3 × 5
# Groups:   med_legal [2]
  med_legal rec_legal mean_all     n mean_drivers
  <lgl>     <lgl>        <dbl> <int>        <dbl>
1 FALSE     FALSE         15.1   653         15.5
2 TRUE      FALSE         17.8   361         17.7
3 TRUE      TRUE          22.3   108         21.0
```

Our summary statistics are immediately interesting. First, we see that pay across all occupations and for drivers specifically is higher when marijuana is legal. This makes sense, since thus far marijuana legalization has been a one-way ratchet, and nominal wages tend to climb with time. More interesting than that, we notice that when marijuana is illegal in a given state-year, the mean of median wages for drivers are higher than for all other occupations, while wages for drivers are actually lower than all other occupations in legal state-years. This could be evidence that marijuana legalization actually predicts lower wages, but there are other reasonable explanations. It's virtually certain that the states that have legalized marijuana are not a random subset of states, and it's possible that these states are states in which drivers command relatively low wages. We'll have to consider the impact of legalization in each state specifically to account for this in our final statistical tests. Finally, it's just interesting to note that despite grouping by two binary variable, we have only three cases, because there are no state-years in which recreational marijuana is legal but medicinal marijuana is illegal.