Rent in San Francisco

TidyTuesday Challenge (from 7/5/22)

Alan Wang

INSERT DATE

Set Up

```
# load packages
library(tidyverse)
library(tidymodels)
library(knitr)
library(dplyr)
library(ggmap)

# load data
rent <- readr::read_csv('https://raw.githubusercontent.com/rfordatascience/tidytuesday/mas
```

Overview of Data

This data were collected by Dr. Kate Pennington, who scraped Craiglist rental postings of properties in the general San Francisco area from 2000 to 2018. The dataset contains 200,796 observations, each of which represents an individual unit, identified by a unique ID, Craiglist posting date, neighborhood, city, county, monthly rent price, number of beds/baths, size in terms of square feet, whether there is room in the apartment, the address, and a description.

Data Preparation

The following report of the data will primarily focus on the rent price, city, county, number of beds, number of baths, size, year, and whether there is room in the apartment. Thus, before any visualizations are made or any analysis is conducted, any observations that do not have values for the aforementioned variables will be dropped. The `room_in_apt` variable will also be made into a factor variable for regression analysis. This new dataset has 14,629 observations.

```r
rents <- rent |>
  drop_na(year | city | county | price | beds | baths | sqft | room_in_apt) |>
  mutate(room_in_apt = as.factor(room_in_apt))
```

Linear Regression

Creating the Model

This is a linear regression model that will be used to predict monthly rent price for rental units in San Francisco with year posted on Craiglist, number of beds, number of paths, number of square feet, and whether or not there is room in the apartment.

```r
set.seed(1)

rents_split <- initial_split(data = rents)
rents_train <- training(rents_split)
rents_test <- testing(rents_split)
```

```r
rents_spec <- linear_reg() |>
  set_engine("lm")
```

```r
rents_rec <- recipe(price ~ year + beds + baths + sqft + room_in_apt, data = rents) |>
  step_center(all_numeric_predictors()) |>
  step_dummy(all_nominal_predictors()) |>
  step_zv(all_predictors())
```

```r
rents_wflow <- workflow() |>
  add_model(rents_spec) |>
  add_recipe(rents_rec)
```

```r
rents_fit <- rents_wflow |>
  fit(data = rents_train)

rents_fit |>
  tidy() |>
  kable(digits = 3)
```

term	estimate	std.error	statistic	p.value
(Intercept)	2795.709	12.098	231.080	0.000
year	183.804	4.830	38.051	0.000
beds	5.537	18.023	0.307	0.759
baths	321.814	24.007	13.405	0.000
sqft	1.018	0.027	37.448	0.000
room_in_apt_X1	-468.117	234.590	-1.995	0.046

For every one additional year, we expect that the average monthly rent price will increase by $183.80, on average, holding all else constant.

For every one additional bedroom in the rental property, we expect that the average monthly rent price will increase by $5.54, on average, holding all else constant.

For every one additional bathroom in the rental property, we expect that the average monthly rent price will increase by $321.81, on average, holding all else constant.

For every one additional square foot of the rental property size, we expect that the average monthly rent price will increase by $1.02, on average, holding all else constant.

We expect that a rental apartment that has space available will have an average monthly rent price that is $468.12 less than a rental apartment that does not have space available, on average, holding all else constant.

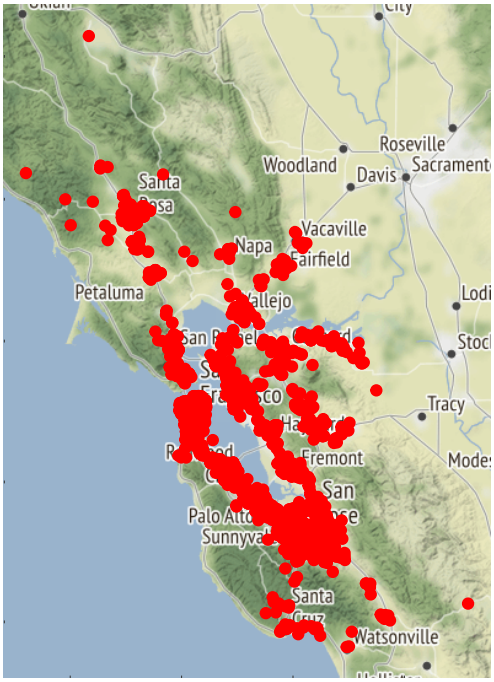
Visualizations

Map of Rental Properties by Location

This visualization displays where properties are located by longitude and latitude. The data was filtered to only include valid longitude and latitude coordinates from the previously filtered `rents` dataset.

```
rent_maps <- rents |>
  drop_na(lat | lon)
```

```
qmplot(lon, lat, data = rent_maps, maptype = "toner-lite", color = I("red"), zoom = 8)
```

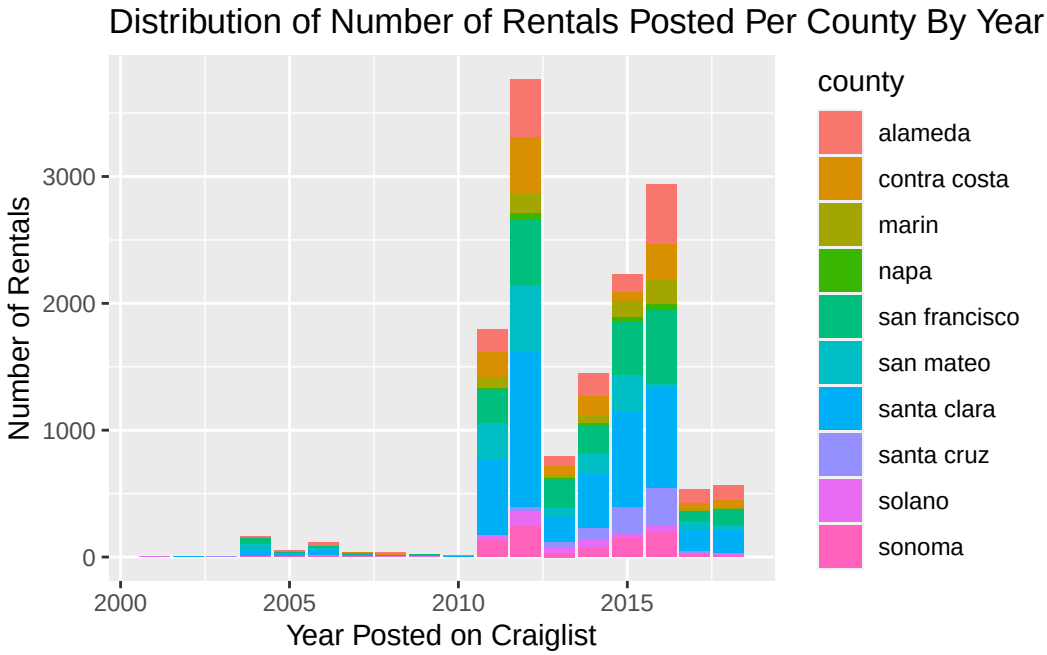


The map shows that the majority of the rental properties are clustered around the San Francisco Bay area, with a very high concentration in the city of San Francisco, as well as in the southern region of the Bay Area near Sunnyvale and San Jose. There also appear to be several outliers that are further north of the San Francisco Bay.

Number of Rentals Posted Per County By Year

This visualization displays

```
ggplot(data = rents, aes(fill = county, x = year)) +
  geom_bar() +
  labs(x = "Year Posted on Craiglist",
       y = "Number of Rentals",
       title = "Distribution of Number of Rentals Posted Per County By Year")
```



The plot shows that the number of rentals posted was the highest in 2011, and the number of rentals posted appears to significantly increase from 2010 to 2011, with very little postings between 2000 and 2010. There also appears to be sharp drops in the number of rental postings between 2011 and 2012 and 2016 and 2017. Between 2011 and 2018, Santa Clara County consistently has the highest number of rentals posted compared to other counties.