Bikeshare Dataset

1 Introduction

This dataset includes daily counts of rented bicycles from the Washington, D.C.-based bicycle rental firm Capital-Bikeshare, as well as weather and seasonal data. Capital-Bikeshare generously made the data publicly available. Fanaee-T and Gama (2014) incorporated weather and seasonal data. The objective is to forecast how many bikes will be booked based on the weather and the time of day. For more details regarding the dataset, please refer to Dua and Graff (2017).

Figure 1: A bikesharing service (Internet image)

Dataset basic information:

Variable	Description
`cnt` (target)	count of total rental bikes
`weathersit`	weather situation (GOOD, or MISTY, or RAIN/SNOW/STORM)
`temp`	temperature in Celsius.
`hum`	humidity in percent
`windspeed`	wind speed in km/h
`season`	season (WINTER, SPRING, SUMMER, FALL)
`yr`	year (2011, 2012)
`mnth`	month of year (JAN, FEB, ..., DEC)
`weekday`	day of the week (SUN, MON, ..., SAT)
`holiday`	indicator whether it is a holiday or not
`workingday`	YES if day is not weekend, otherwise is NO.

```
# load the dataset from OpenML Library
d <- OpenML::getOMLDataSet(data.id = 41979)
# convert the OpenML object to a tibble (enhanced data.frame)
bikeshare <- d %>% dplyr::as_tibble()
skimmed_bikeshare <- skimr::skim(bikeshare)
print(bikeshare, width = Inf)
```

```
## # A tibble: 731 x 11
##    season yr    mnth  holiday    weekday workingday     weathersit   temp   hum
##    <fct>  <fct> <fct> <fct>      <fct>   <fct>          <fct>       <dbl> <dbl>
##  1 SPRING 2011  JAN   NO HOLIDAY SAT     NO WORKING DAY MISTY       8.18   80.6
##  2 SPRING 2011  JAN   NO HOLIDAY SUN     NO WORKING DAY MISTY       9.08   69.6
##  3 SPRING 2011  JAN   NO HOLIDAY MON     WORKING DAY    GOOD        1.23   43.7
##  4 SPRING 2011  JAN   NO HOLIDAY TUE     WORKING DAY    GOOD        1.4    59.0
##  5 SPRING 2011  JAN   NO HOLIDAY WED     WORKING DAY    GOOD        2.67   43.7
##  6 SPRING 2011  JAN   NO HOLIDAY THU     WORKING DAY    GOOD        1.60   51.8
##  7 SPRING 2011  JAN   NO HOLIDAY FRI     WORKING DAY    MISTY       1.24   49.9
##  8 SPRING 2011  JAN   NO HOLIDAY SAT     NO WORKING DAY MISTY      -0.245  53.6
##  9 SPRING 2011  JAN   NO HOLIDAY SUN     NO WORKING DAY GOOD       -1.50   43.4
## 10 SPRING 2011  JAN   NO HOLIDAY MON     WORKING DAY    GOOD       -0.911  48.3
##    windspeed   cnt
##        <dbl> <dbl>
##  1     10.7    985
##  2     16.7    801
##  3     16.6   1349
##  4     10.7   1562
##  5     12.5   1600
##  6      6.00  1606
##  7     11.3   1510
##  8     17.9    959
##  9     24.3    822
## 10     15.0   1321
## # ... with 721 more rows
```

2 Exploratory Data Analysis (EDA)

In this part, we will walk through a few characteristics of bikeshare dataset using library `skimr` and `DataExplorer`.

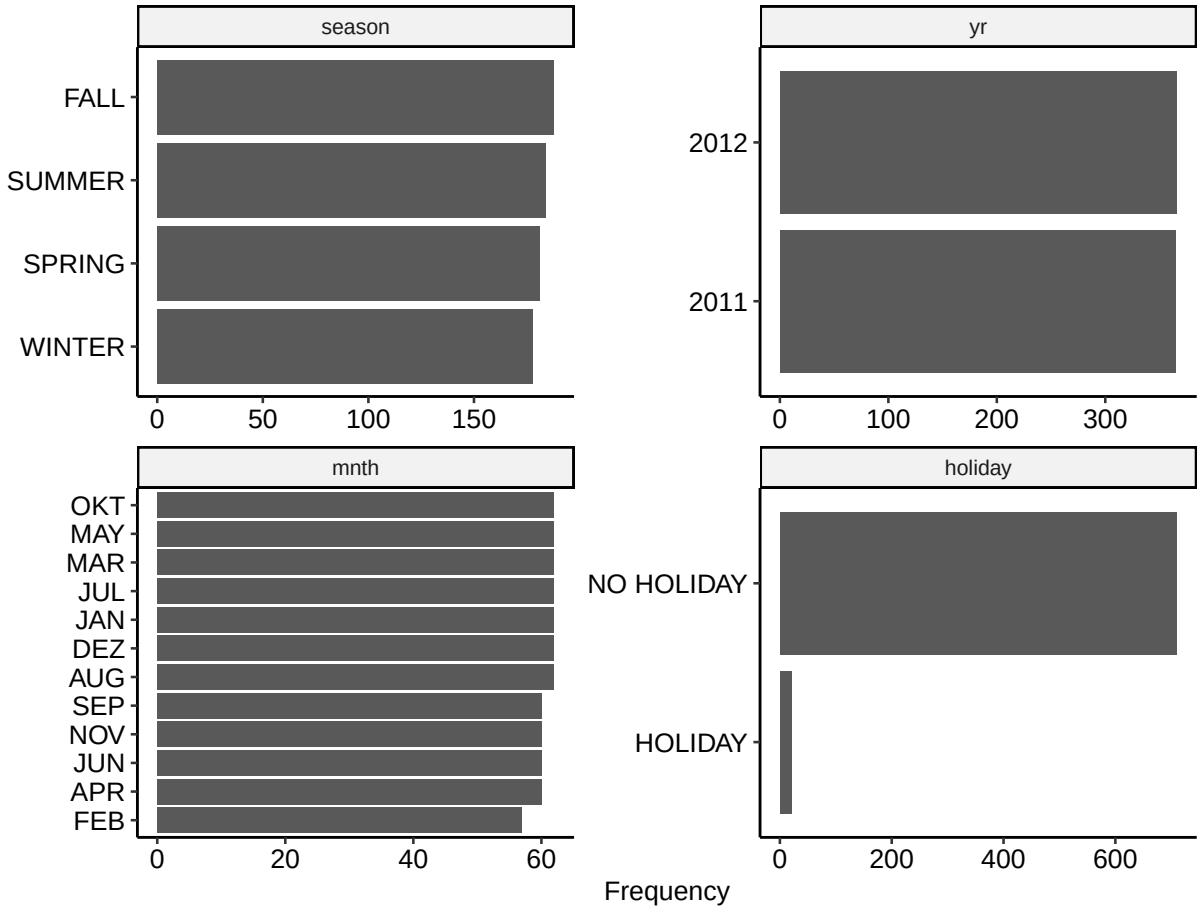
2.1 Factor variables

General statistics about factor variables from bikeshare dataset:

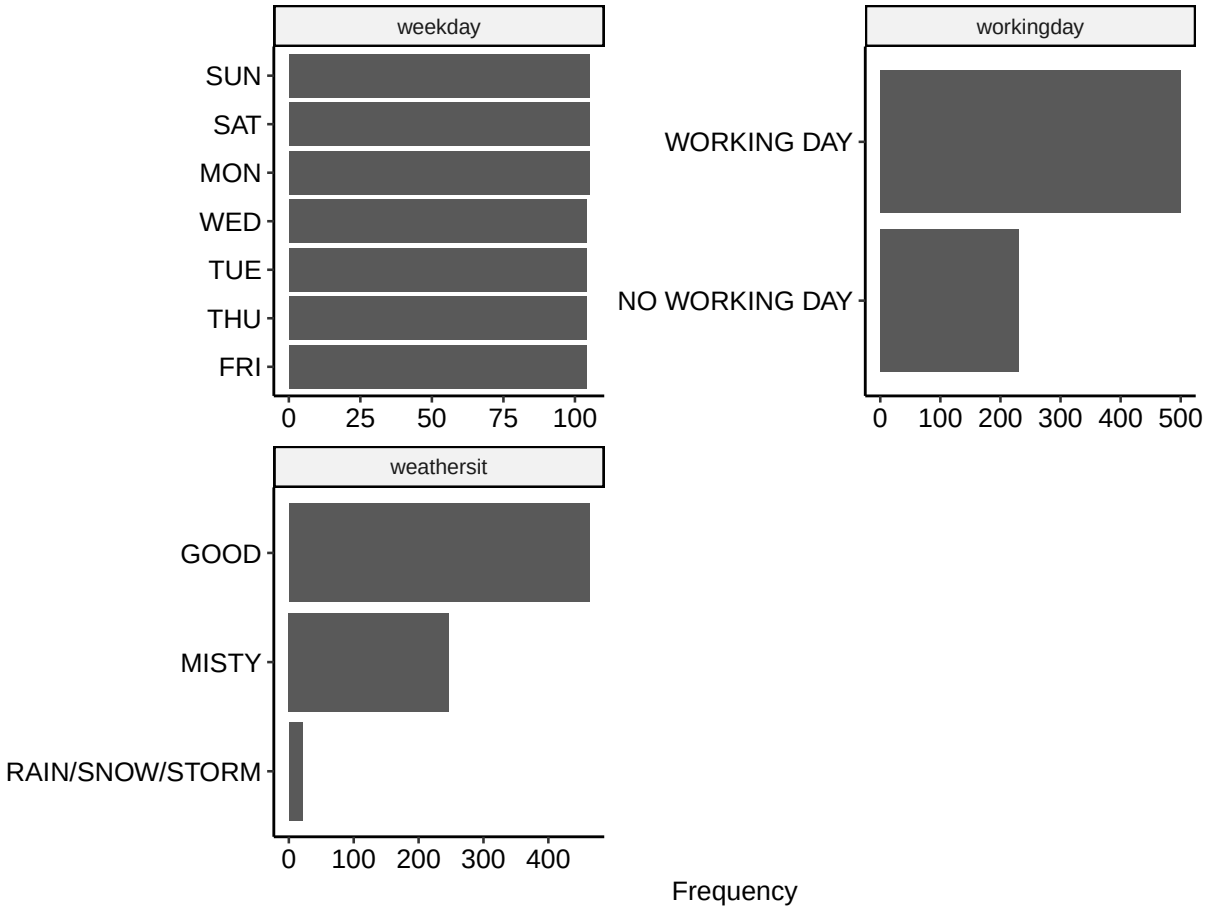
```
skimr::partition(skimmed_bikeshare)$factor %>%
        knitr::kable(format = 'latex', booktabs = TRUE) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
season	0	1	FALSE	4	FAL: 188, SUM: 184, SPR: 181, WIN: 178
yr	0	1	FALSE	2	201: 366, 201: 365
mnth	0	1	FALSE	12	JAN: 62, MAR: 62, MAY: 62, JUL: 62
holiday	0	1	FALSE	2	NO : 710, HOL: 21
weekday	0	1	FALSE	7	SUN: 105, MON: 105, SAT: 105, TUE: 104
workingday	0	1	FALSE	2	WOR: 500, NO : 231
weathersit	0	1	FALSE	3	GOO: 463, MIS: 247, RAI: 21

```
DataExplorer::plot_bar(
        bikeshare,
        ggtheme = ggpubr::theme_pubr(base_size = 10),
        ncol = 2,
        nrow = 2
)
```



Page 1



This dataset contains 7 factor variables: `season`, `yr`, `mnth`, `holiday`, `weekday`, and `weathersit`. There is no missing data in these variables. With these 7 features, `season`, `yr`, `mnth` and `weekday` have balanced distribution across their categories. Most instances of the dataset are from non-holiday days (accounting for 97.13% of the sample size). There are more instances from working days than non-working days (68.4% of the number of instances). The weather situation is mostly good, followed by misty and lastly rain/snow/storm, with the respective percentages 63.3%, 33.8%, and 2.9%.

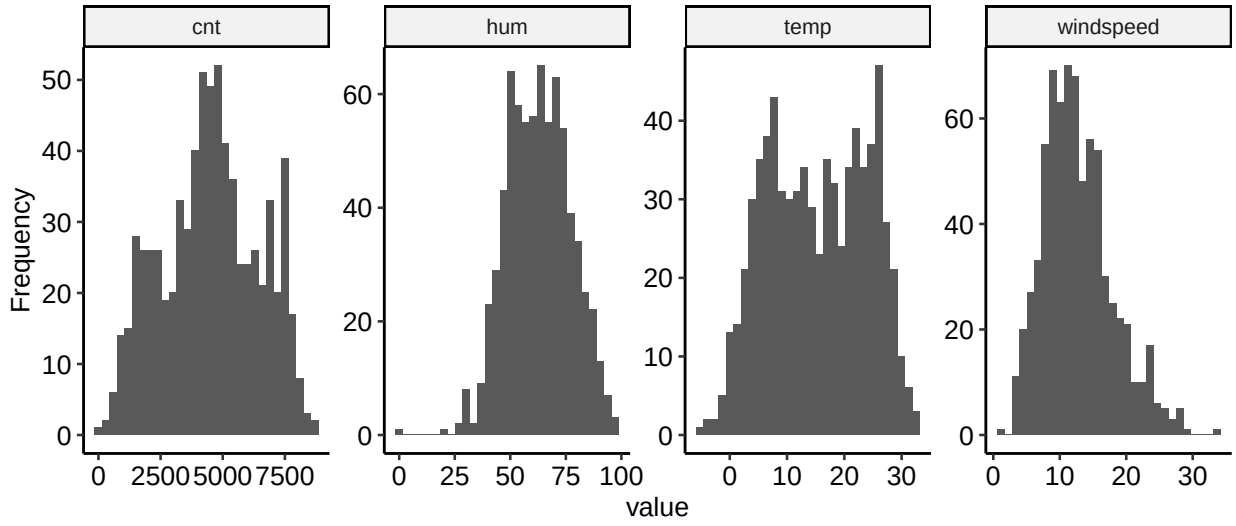
2.2 Numerical variables

General statistics about numerical variables from bikeshare dataset:

```
skimr::partition(skimmed_bikeshare)$numeric %>%
        knitr::kable(format = 'latex', booktabs = TRUE, digits = 2) %>%
        kableExtra::kable_styling(latex_options = 'HOLD_position')
```

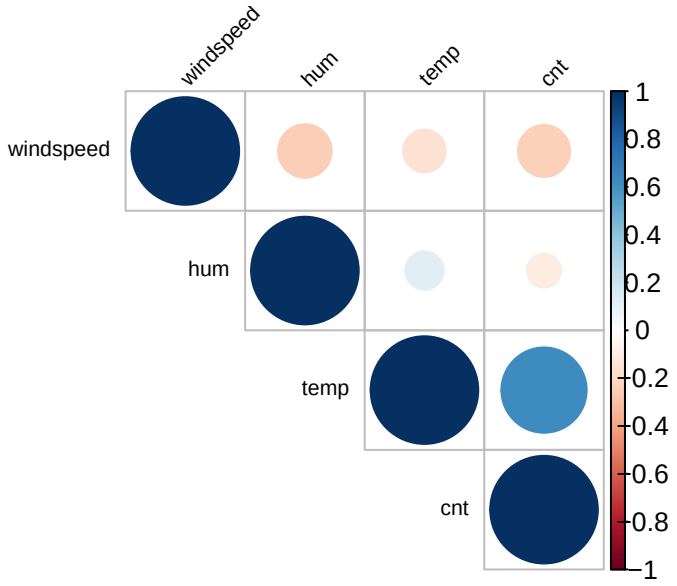
skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
temp	0	1	15.28	8.60	-5.22	7.84	15.42	22.80	32.50	
hum	0	1	62.79	14.24	0.00	52.00	62.67	73.02	97.25	
windspeed	0	1	12.76	5.19	1.50	9.04	12.13	15.63	34.00	
cnt	0	1	4504.35	1937.21	22.00	3152.00	4548.00	5956.00	8714.00	

```
DataExplorer::plot_histogram(
        bikeshare,
        ggtheme = ggpubr::theme_pubr(base_size = 10)
)
```



Similar to the factor variables, there is no missing value in the numerical variables. `cnt` and `hum` have a roughly symmetric distribution, in which `hum`'s distribution also seems to have a bell shape. `temp` appears to have a bimodal distribution with the peaks at 10 and 25. Lastly, `windspeed`'s distribution is slightly right skewed.

```
bikeshare_numeric <- bikeshare %>% select(where(is.numeric))
bikeshare_numeric %>%
        cor() %>%
        corrplot(
           type = "upper",
           order = "hclust",
           tl.col = "black",
           tl.srt = 45,
           tl.cex = 0.7
        )
```



Through the correlation plot above, it is notable that the feature `temp` appears to have a strong positive correlation with the target `cnt`. This can be an indicator that temperature may affect the decision of customers using bikeshare services and that `temp` can be a good candidate for predicting the target.

References

Dua, Dheeru, and Casey Graff. 2017. "UCI Machine Learning Repository." University of California, Irvine, School of Information; Computer Sciences. http://archive.ics.uci.edu/ml.

Fanaee-T, Hadi, and João Gama. 2014. "Event Labeling Combining Ensemble Detectors and Background Knowledge." *Progress in Artificial Intelligence* 2 (June): 113–27. https://doi.org/10.1007/s13748-013-0040-3.