Wybrany zbiór danych: Seoul Bike Sharing Demand Data Set

https://archive.ics.uci.edu/ml/datasets/Seoul+Bike+Sharing+Demand

– zawiera ilość wypożyczonych rowerów w godzinę w wypożyczalni rowerów w Seoulu, w zależności od warunków pogodowych, pór roku i dni świątecznych. Zbiór pochodzi z oficjalnej strony internetowej Seoulu. Na stronie UCI został opublikowany 2020-03-01. Zgodnie z opisem zawiera 8760 rekordów i 14 atrybutów, które przyjmują następujące wartości:

Date : (rok-miesiąc-dzień)
Rented Bike Count – Ilość rowerów wypożyczonych w każdej godzinie
Hour – Obecna godzina
Temperature-Temperatura w stopniach Celsjusza
Humidity – Wilgotność podana w procentach
Windspeed -Szybkość wiatru w m/s
Visibility – Widoczność w metrach
Dew point temperature – Temperatura punktu rosy (temperatura, w której może rozpocząć się proces skraplania gazu lub wybranego składnika mieszaniny gazów przy ustalonym ciśnieniu, a w przypadku mieszaniny gazów również przy określonym składzie) w stopniach Celsjusza
Solar radiation – Promieniowanie słoneczne w MJ/m2
Rainfall – Opad deszczu w mm
Snowfall – Opad śniegu w cm
Seasons - Winter, Spring, Summer, Autumn(pora roku)
Holiday - Holiday/No holiday(dzień świąteczny czy powszedni)
Functional Day – czy w danym dniu wypożyczalnia była czynna - NoFunc(Non Functional Hours), Fun(Functional hours).

Tworzymy ramkę danych z naszego zbioru, sprawdzamy nazwy kolumn oraz ich typy zmiennych:

```
> print(names(df))
 [1] "Date"                        "Rented.Bike.Count"
 [3] "Hour"                        "Temperature..C."
 [5] "Humidity..."                 "Wind.speed..m.s."
 [7] "Visibility..10m."            "Dew.point.temperature..C."
 [9] "Solar.Radiation..MJ.m2."     "Rainfall.mm."
[11] "Snowfall..cm."               "Seasons"
[13] "Holiday"                     "Functioning.Day"
```

Tabela 1. Nazwy kolumn

```
> str(df)
'data.frame':   8760 obs. of  14 variables:
 $ Date                     : chr  "01/12/2017" "01/12/2017" "01/12/2017" "01/12/2017" ...
 $ Rented.Bike.Count        : int  254 204 173 107 78 100 181 460 930 490 ...
 $ Hour                     : int  0 1 2 3 4 5 6 7 8 9 ...
 $ Temperature..C.          : num  -5.2 -5.5 -6 -6.2 -6 -6.4 -6.6 -7.4 -7.6 -6.5 ...
 $ Humidity...              : int  37 38 39 40 36 37 35 38 37 27 ...
 $ Wind.speed..m.s.         : num  2.2 0.8 1 0.9 2.3 1.5 1.3 0.9 1.1 0.5 ...
 $ Visibility..10m.         : int  2000 2000 2000 2000 2000 2000 2000 2000 2000 1928 ...
 $ Dew.point.temperature..C.: num  -17.6 -17.6 -17.7 -17.6 -18.6 -18.7 -19.5 -19.3 -19.8 -22.4 ...
 $ Solar.Radiation..MJ.m2.  : num  0 0 0 0 0 0 0 0 0.01 0.23 ...
 $ Rainfall.mm.             : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Snowfall..cm.            : num  0 0 0 0 0 0 0 0 0 0 ...
 $ Seasons                  : chr  "Winter" "Winter" "Winter" "Winter" ...
 $ Holiday                  : chr  "No Holiday" "No Holiday" "No Holiday" "No Holiday" ...
 $ Functioning.Day          : chr  "Yes" "Yes" "Yes" "Yes" ...
```

Tabela 2. Opis danych

Możemy zauważyć, że opis danych jest nie do końca zgodny z rzeczywistością, gdyż atrybut Functioning.Day posiada wartości ‘Yes’. Po sprawdzeniu:

```
> unique(df$Functioning.Day)
[1] "Yes" "No"
```

Tabela 3. Wartości atrybutu Functioning.Day

Atrybut ten mówi nam czy w danym dniu funkcjonuje wypożyczalnia. Dla pewności sprawdziliśmy średnią dla dni w których wypożyczalnia była zamknięta i faktycznie średnia wypożyczonych rowerów wynosiła 0. Możemy więc rozważyć usunięcie tej kolumny.

Analiza jednowymiarowa

Statystyka opisowa

```
[1] "Date                      : 8760"
[1] "Rented.Bike.Count         : 8760"
[1] "Hour                      : 8760"
[1] "Temperature..C.           : 8760"
[1] "Humidity...               : 8760"
[1] "Wind.speed..m.s.          : 8760"
[1] "Visibility..10m.          : 8760"
[1] "Dew.point.temperature..C.: 8760"
[1] "Solar.Radiation..MJ.m2.   : 8760"
[1] "Rainfall.mm.              : 8760"
[1] "Snowfall..cm.             : 8760"
[1] "Seasons                   : 8760"
[1] "Holiday                   : 8760"
[1] "Functioning.Day           : 8760"
[1] "Bike.Count.Group          : 8760"
```

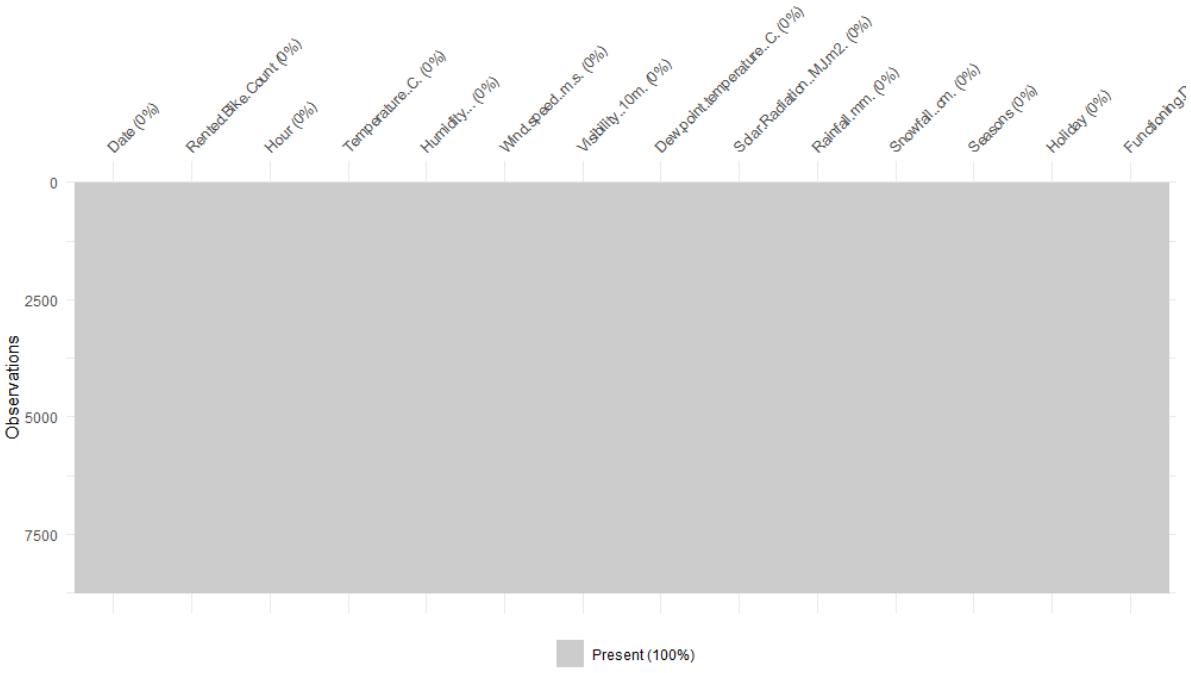
Tabela 4. Liczba obserwacji z wyłączeniem brakujących

```
[1] "Date                      : 0"
[1] "Rented.Bike.Count         : 0"
[1] "Hour                      : 0"
[1] "Temperature..C.           : 0"
[1] "Humidity...               : 0"
[1] "Wind.speed..m.s.          : 0"
[1] "Visibility..10m.          : 0"
[1] "Dew.point.temperature..C.: 0"
[1] "Solar.Radiation..MJ.m2.   : 0"
[1] "Rainfall.mm.              : 0"
[1] "Snowfall..cm.             : 0"
[1] "Seasons                   : 0"
[1] "Holiday                   : 0"
[1] "Functioning.Day           : 0"
[1] "Bike.Count.Group          : 0"
```

Tabela 5. Liczba brakujących

Możemy zauważyć, że w naszym zbiorze nie ma brakujących wartości

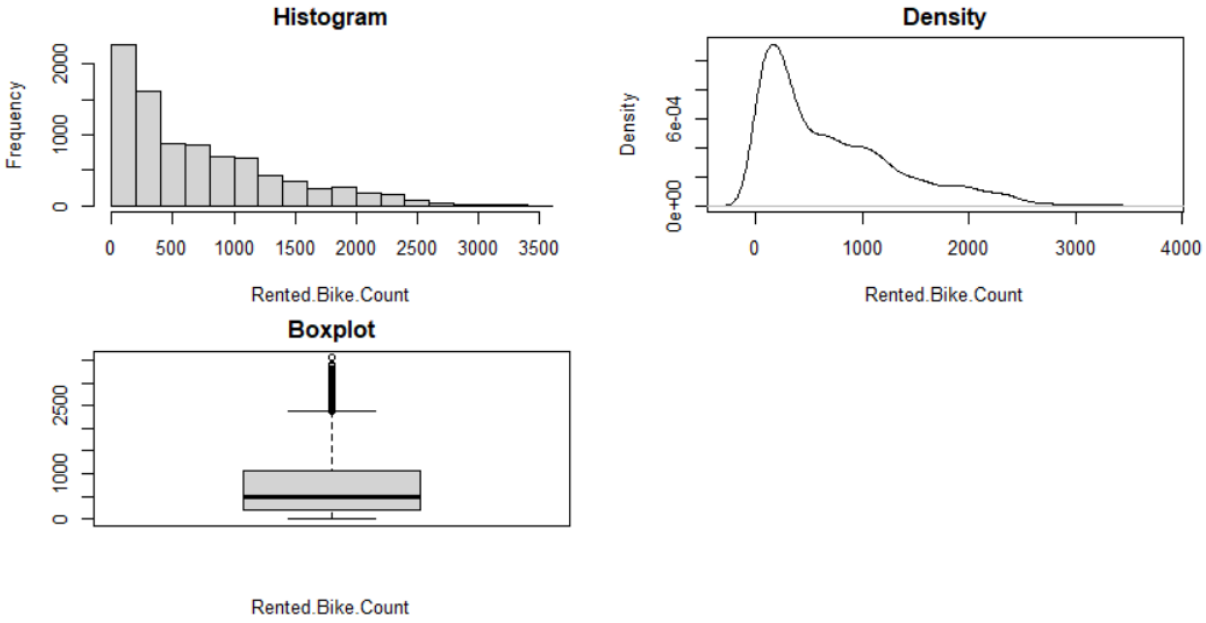
Dodatkowo potwierdziliśmy to funkcją vis_miss z biblioteki naniar:



Rys. 1. Wykres brakujących wartości w poszczególnych kolumnach

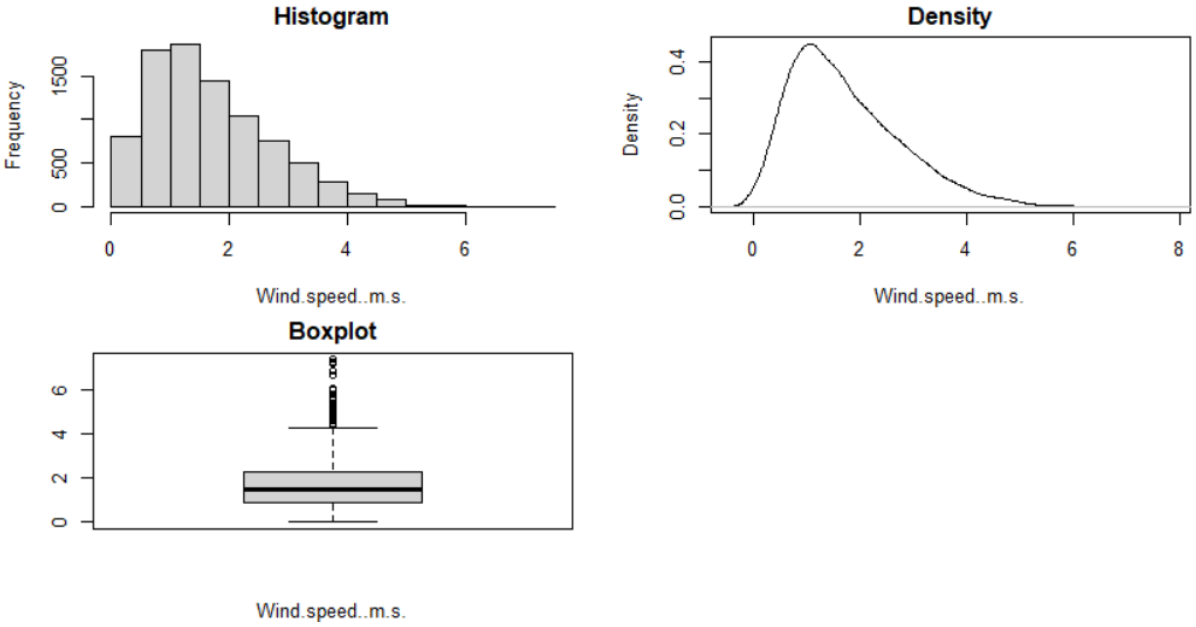
Okazało się, że zbiór jest kompletny.

Następnie wybraliśmy najciekawsze wykresy dotyczące jednowymiarowych danych:



Rys. 2.1. Wykresy (Histogram, Density, Boxplot) jednowymiarowe dla ilości wypożyczonych rowerów

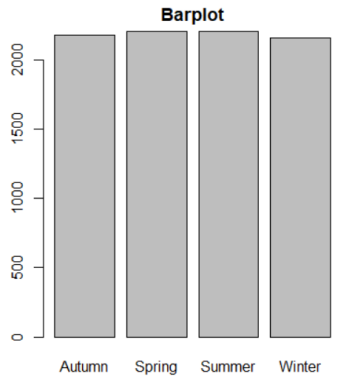
Powyższe wykresy dotyczą ilości wypożyczonych rowerów w jednej godzinie. Widzimy, że wykres gęstości ma charakterystykę podobną do wykładniczej, a na histogramie zdecydowanie najczęściej występuje ilość wypożyczeń w zakresie 0-200 oraz trochę rzadziej w zakresie 200-400. Może to być związane z brakiem wypożyczeń w nocy i okresie zimowym. Na boxplocie możemy zaobserwować, iż tendencja centralna w postaci mediany jest przesunięta w stronę pierwszego kwartyla. Oznacza to, że dane po lewej stronie mediany są gęściej rozłożone niż po prawej stronie. Widzimy także wartości odstające za 3 kwartylem.



Rys. 2.2. Wykresy (Histogram, Density, Boxplot) jednowymiarowe dla prędkości wiatru

Tutaj widzimy wykresy dotyczące prędkości wiatru. Wykres gęstości jest prawoskośny, czyli jego moda jest mniejsza od mediany i średniej arytmetycznej. Możemy to także dobrze zaobserwować na boxplocie.

Następnie sprawdziliśmy wykresy dotyczące danych kategorycznych.



Rys. 3. Wykres (Barplot) jednowymiarowy dla pór roku

Z powyższego wykresu widzimy, że obserwacje były prowadzone przez cały rok, żadna pora roku nie jest dominująca.

Miary położenia i rozrzutu

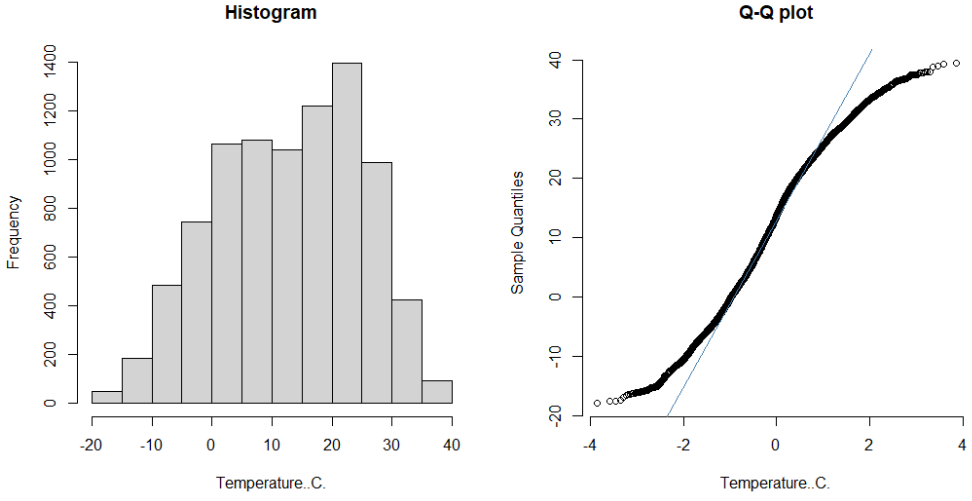
Rented.Bike.Count	Hour	Temperature..C.	Humidity...	Wind.speed..m.s.	Visibility..10m.	Dew.point.temperature..C.	Solar.Radiation..MJ.m2.	Rainfall.mm.	Snowfall..cm.
Min. : 0.0	Min. : 0.00	Min. :-17.80	Min. : 0.00	Min. :0.000	Min. : 27	Min. :-30.600	Min. :0.0000	Min. : 0.0000	Min. :0.00000
1st Qu.: 191.0	1st Qu.: 5.75	1st Qu.: 3.50	1st Qu.:42.00	1st Qu.:0.900	1st Qu.: 940	1st Qu.: -4.700	1st Qu.:0.0000	1st Qu.: 0.0000	1st Qu.:0.00000
Median : 504.5	Median :11.50	Median : 13.70	Median :57.00	Median :1.500	Median :1698	Median : 5.100	Median :0.0100	Median : 0.0000	Median :0.00000
Mean : 704.6	Mean :11.50	Mean : 12.88	Mean :58.23	Mean :1.725	Mean :1437	Mean : 4.074	Mean :0.5691	Mean : 0.1487	Mean :0.07507
3rd Qu.:1065.2	3rd Qu.:17.25	3rd Qu.: 22.50	3rd Qu.:74.00	3rd Qu.:2.300	3rd Qu.:2000	3rd Qu.: 14.800	3rd Qu.:0.9300	3rd Qu.: 0.0000	3rd Qu.:0.00000
Max. :3556.0	Max. :23.00	Max. : 39.40	Max. :98.00	Max. :7.400	Max. :2000	Max. : 27.200	Max. :3.5200	Max. :35.0000	Max. :8.80000
Std: 644.997468	Std: 6.922582	Std: 11.944825	Std: 20.362413	Std: 1.036300	Std: 608.298712	Std: 13.060369	Std: 0.868746	Std: 1.128193	Std: 0.436746
Variance: 416021.733390	Variance: 47.922137	Variance: 142.678850	Variance: 414.627875	Variance: 1.073918	Variance: 370027.323001	Variance: 170.573247	Variance: 0.754720	Variance: 1.272819	Variance: 0.190747
Skewness: 1.153033	Skewness: 0.000000	Skewness: -0.198258	Skewness: 0.059559	Skewness: 0.890650	Skewness: -0.701546	Skewness: -0.367173	Skewness: 1.503525	Skewness: 14.528255	Skewness: 8.437910
Kurtosis: 0.851336	Kurtosis: -1.204584	Kurtosis: -0.838487	Kurtosis: -0.804287	Kurtosis: 0.725229	Kurtosis: -0.962581	Kurtosis: -0.756196	Kurtosis: 1.124164	Kurtosis: 284.762063	Kurtosis: 93.727019

Tabela 6. Miary położenia i rozrzutu

Powyższe dane dokładniej pokazują rozkład danych w atrybutach niż w formie graficznej, jednak wnioski z nich wyciągnięte będą podobne(jednak w formie graficznej łatwiej zauważyć niektóre zależności). Widzimy, że rozkłady nie są mezokurtyczne, gdyż ich kurtoza jest różna od zera. Okazało się, że występują rozkłady zarówno platokurtyczne jak i leptokurtyczne.

Badanie normalności

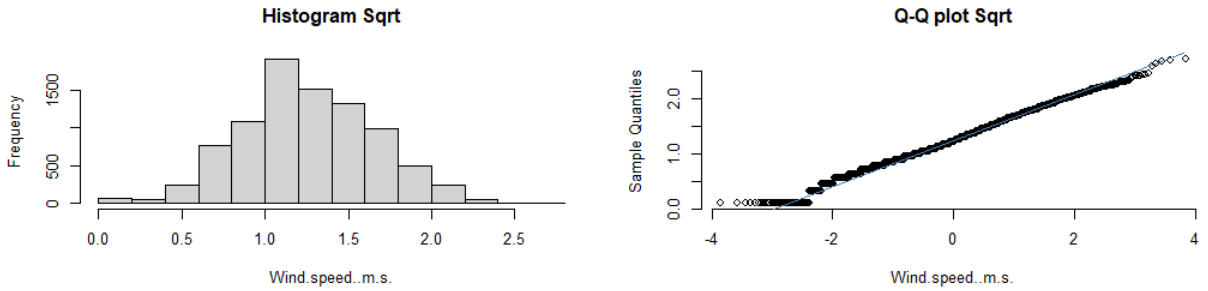
Po przejrzeniu wykresów widzimy, że dane raczej nie rozkładają się zgodnie z rozkładem naturalnym, w naszej ocenie najbardziej do niego zbliżony jest wykres temperatury:



Rys 4.1. Wykresy (Histogram, Q-Q plot) dla temperatury

Aby spróbować uzyskać bardziej znormalizowane dane zastosowaliśmy przekształcenia za pomocą logarytmu i pierwiastka kwadratowego.

Po sprawdzeniu wykresów przekształconych danych, dalej większość z nich nie przypominała rozkładu normalnego z wyjątkiem wykresu prędkości wiatru.



Rys. 4.2. Wykresy (Histogram, Q-Q plot) dla prędkości wiatru

Testy normalności

Wykonaliśmy test Shapiro-Wilka na 30-składnikowych próbkach każdego z atrybutów. W przypadku temperatury, wilgotności, temperatury punktu rosy wykazał on normalność danych(p-value>0.05).

```
[1] "Temperature..C."

        Shapiro-Wilk normality test

data:  x
W = 0.95106, p-value = 0.1804
```

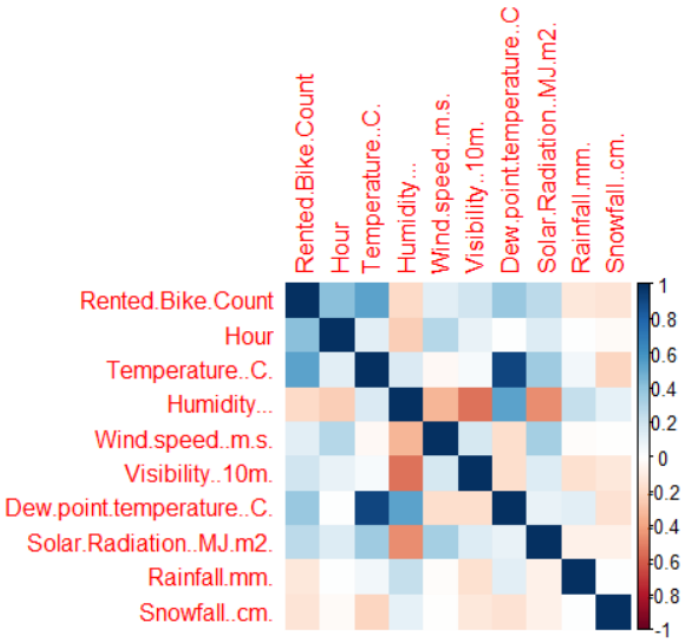
```
[1] "Humidity..."

        Shapiro-Wilk normality test

data:  x
W = 0.98497, p-value = 0.9367
```

```
[1] "Dew.point.temperature..C."

        Shapiro-Wilk normality test

data:  x
W = 0.95855, p-value = 0.2842
```

Tabela 7. Testy normalności

Analiza dwuwymiarowa

Correlation plot



Rys 5. Macierz korelacji pomiędzy wszystkimi atrybutami

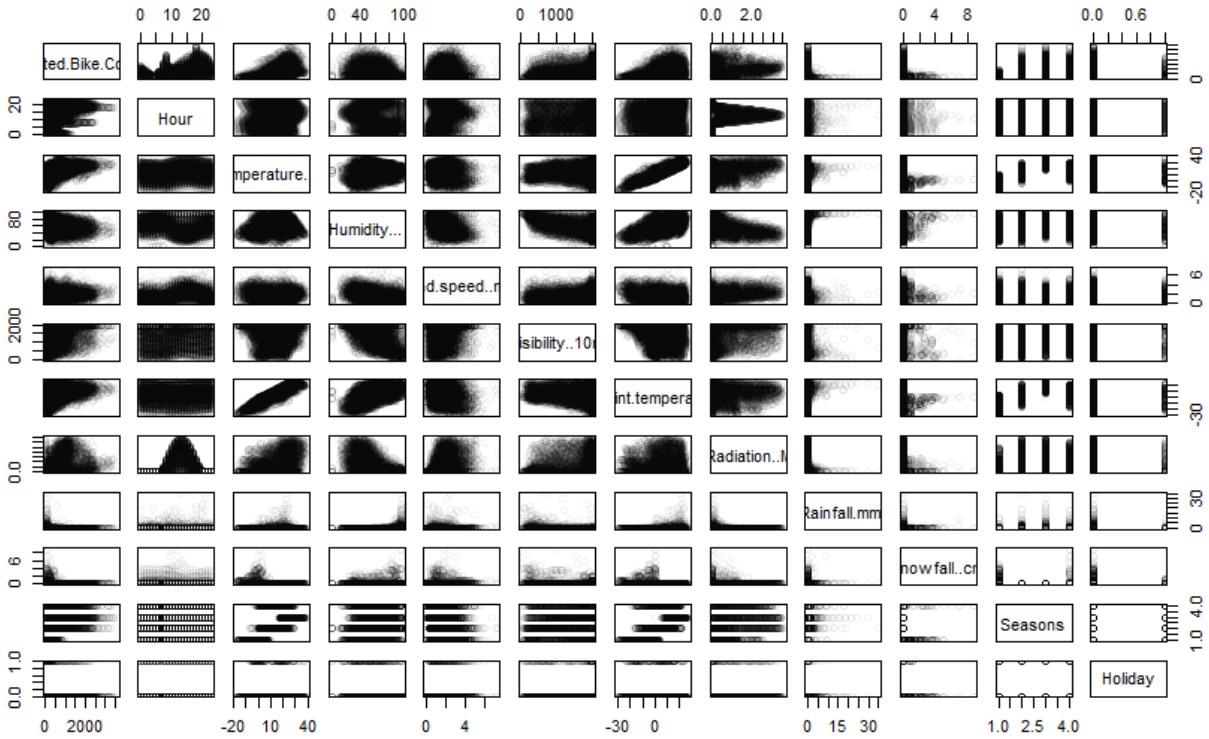
Zależności między atrybutami zostały obliczone za pomocą współczynnika korelacji Pearsona. Oznacza to, że pokazuje on tylko zależności liniowe pomiędzy atrybutami, więc jeżeli pomiędzy atrybutami występuje korelacja wielomianowa, nie zobaczymy jej na tym wykresie.

Możemy zauważyć bardzo dużą korelację pomiędzy temperaturą punktu rosy, a temperaturą powietrza. Ma to odwzorowanie w rzeczywistości, czyli nie jest przypadkowym powiązaniem. W przypadku dalszej analizy moglibyśmy wyeliminować jeden z tych atrybutów. Widzimy także zależności pomiędzy temperaturą a promieniowaniem słonecznym, widocznością a wilgotnością, temperaturą punktu rosy a wilgotnością, promieniowaniem słonecznym a wilgotnością. Są to zależności, które wynikają ze zjawisk fizycznych opisujących pogodą, więc nie są ,,szumem".

Jeśli chodzi o powiązania z najbardziej interesującą nas wartością, czyli ilością wypożyczonych rowerów, największy wpływ mają na nią godzina i temperatura (analogicznie temperatura punktu rosy). Trochę zaskoczył nas niski wpływ poziomu opadów atmosferycznych, cofnęliśmy się jednak do wykresów z analizy jednostkowej i zauważyliśmy, że w Seulu występują one bardzo rzadko, co może być tego przyczyną.

Po zastanowieniu zmieniliśmy niektóre atrybuty kategoryczne na liczbowe. Ta operacja była możliwa, gdyż atrybuty się do tego nadawały (były to pory roku, nadaliśmy im wartości 1,2,3,4, trzeba zaznaczyć, że przy trenowaniu modelu przewidującego ilość wypożyczonych rowerów byłoby to niewskazane ze względu na zwartościowanie pór roku posiadających równorzędne wartości, jednak przy wizualizacji możemy zastosować taki zabieg, oprócz tego zmieniliśmy wartości atrybutu Holiday na 0,1 gdyż oznaczają czy taki dzień występuje czy nie).

Scatterplot matrix



Rys. 6.1. Macierz Scatterplot z zależnościami pomiędzy atrybutami

Z macierzy wykresów widzimy, że rzeczywiście temperatura ma wysoki wpływ na ilość wypożyczonych rowerów. Widzimy, że do pewnego momentu im wyższa temperatura tym więcej wypożyczonych rowerów, jednak przy pewnej temperaturze ilość ta zaczyna spadać, co jest logiczne, gdyż ciężko jest jeździć w wysokiej temperaturze (zależność nieliniowa). Jeśli chodzi o wykres ilości wypożyczonych rowerów względem godziny, możemy zauważyć skoki w godzinach kiedy większość osób przemieszcza się do pracy i z powrotem.

Ponownie widzimy zależności między zjawiskami pogodowymi, które są powiązane ze sobą zjawiskami fizycznymi.

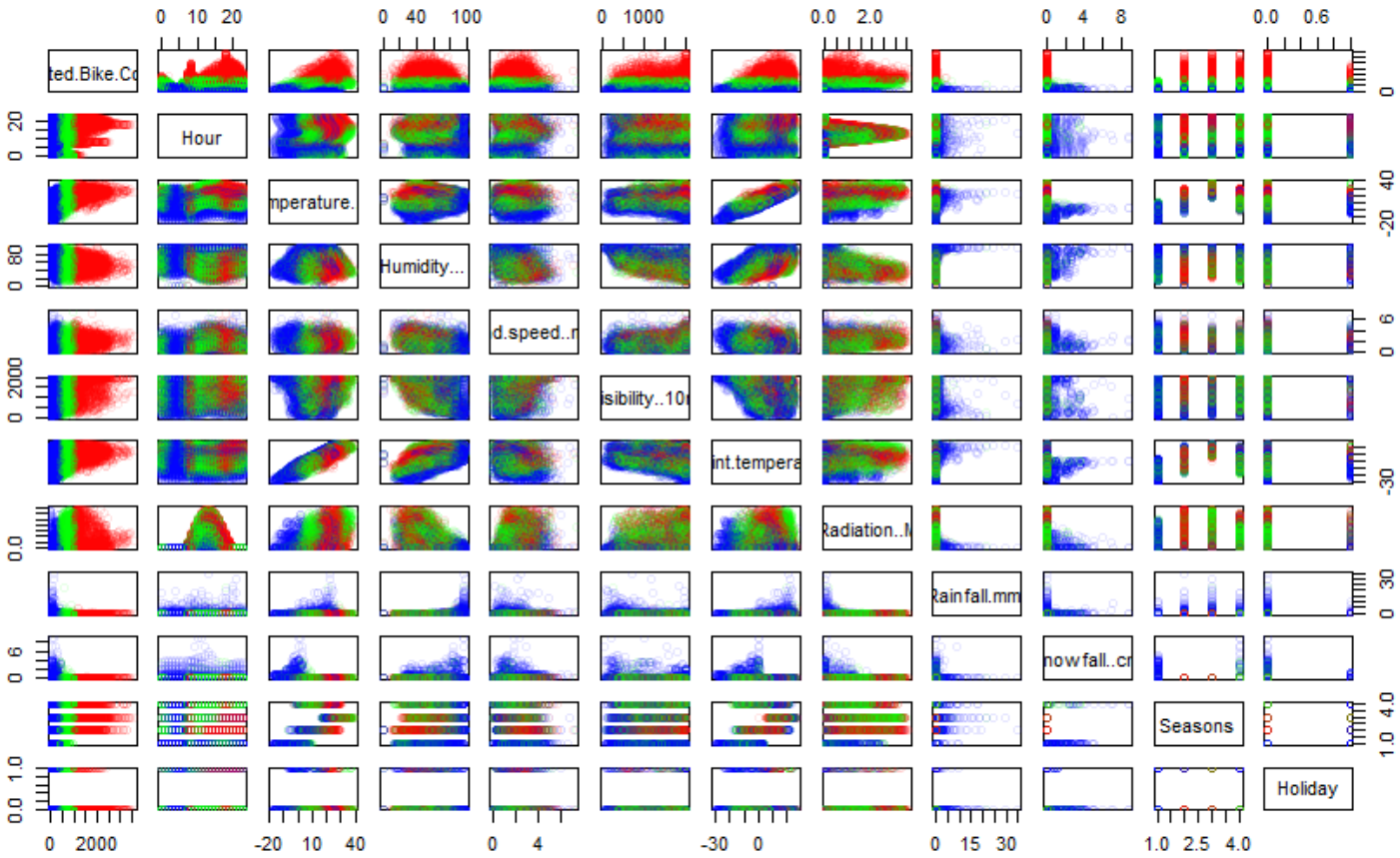
Możemy również zauważyć ciekawe powiązanie pomiędzy ilością wypożyczonych rowerów a widocznością, gdzie przy dużej widoczności maksymalna ilość wypożyczonych rowerów jest wysoka i przy zmianie widoczności gwałtownie

spada w 2 punktach. Oznacza to, że bardzo wysoka widoczność mocno zachęca do jazdy rowerem, z kolei bardzo niska widoczność bardzo zniechęca.

Jeśli chodzi o pory roku widzimy, że wartości atrybutów pogodowych przyjmują wartości spodziewane w danych porach roku np. opady śniegu występują głównie zimą, temperatura jest najwyższa latem itd..

Scatterplot Matrix względem klasy

W związku z tym, że opisywany dataset dotyczy problemu regresyjnego postanowiliśmy dodać nową kolumnę, w której pogrupowaliśmy ilość wypożyczonych rowerów na 3 klasy (‘S’ - small, ‘M’ - medium, ‘L’ - large). Jest to niezbędne do wykonania tego wykresu i jednocześnie przydatne w wizualizacji wpływu nie tylko jednego, lecz większej ilości czynników na liczbę wypożyczonych rowerów.



Rys. 6.2. Macierzy Scatterplot z zależnościami pomiędzy atrybutami względem klasy Bike.Count.Group

‘S’ - kolor niebieski, ‘M’ - kolor zielony, ‘L’ - kolor czerwony

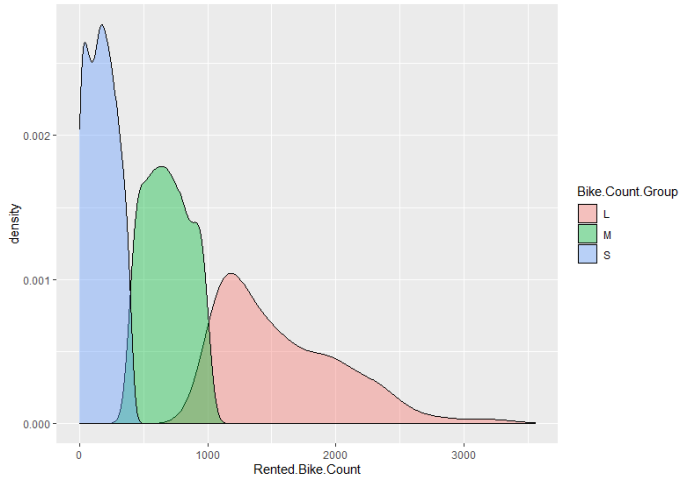
Na wykresach widać, że w pewnych godzinach wypożyczalność rowerów jest niska bez względu na inne atrybuty. Można zauważyć, że w odpowiedniej temperaturze niezależnie od innych warunków atmosferycznych wypożyczalność jest wysoka. Widać, że w przypadku opadów niezależnie od innych warunków rowery nie są wypożyczane.

Dosyć ciekawym wykresem jest wykres Humidity od Dew Point Temperature. Jeśli rośnie wilgotność a temperatura punktu rosy jest odpowiednio niska, wtedy wypożyczalność rowerów jest mała. Są to warunki fizyczne, w których występują mgły, szron, mgła. Są to zjawiska dobrze widoczne dla ludzi, w związku z tym mocno wpływają na odbiór rzeczywistości, w wyniku czego ludzie nie decydują się na wypożyczenie roweru.

Możemy zauważyć, że najmniejsza liczba rowerów była wypożyczana zimą, co jest logiczne. Co ciekawe w dni świąteczne rowery także były wypożyczane, co oznacza, że nie są one używane wyłącznie do dojazdów do pracy.

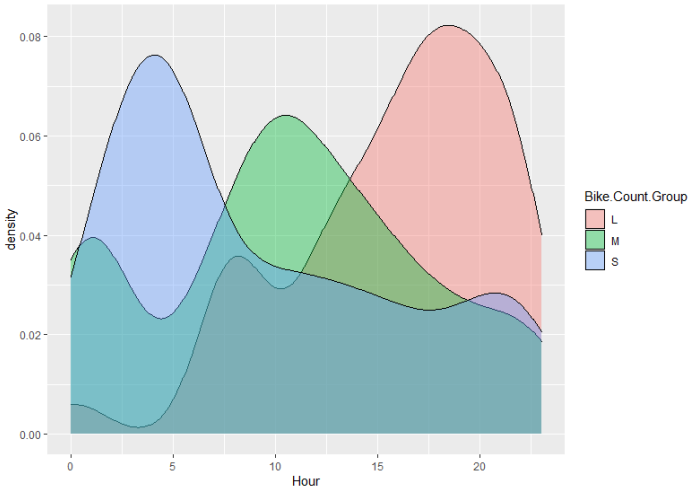
Density względem klasy

Wybraliśmy kilka ciekawych wykresów:



Rys. 7.1. Wykres gęstości dla ilości wypożyczonych rowerów względem klasy Bike.Count.Group

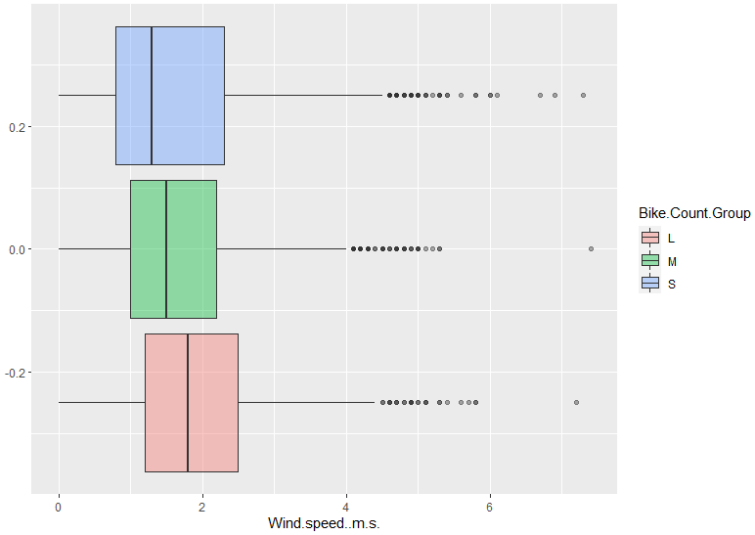
Wykres ten pokazuje odpowiedni dobór wielkości grup ilości wypożyczonych rowerów.



Rys. 7.2. Wykres gęstości dla godzin względem klasy Bike.Count.Group

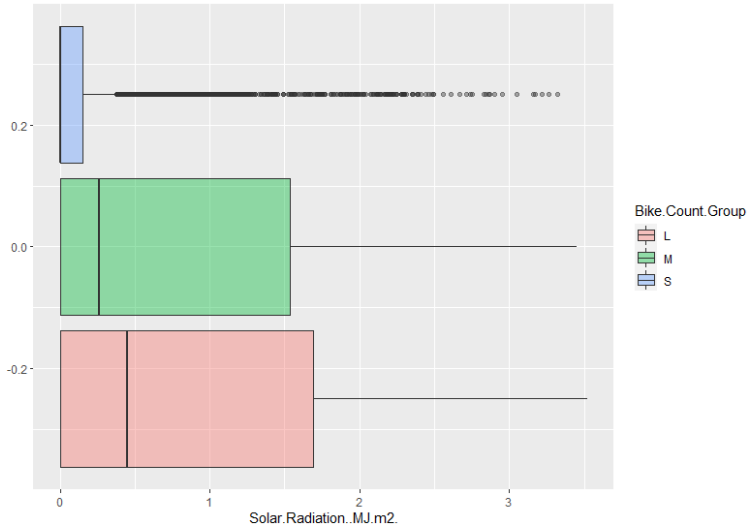
Wykres ten pokazuje jak rozkłada się ilość wypożyczonych rowerów w zależności od godziny. Najmniej rowerów jest wypożyczanych w godzinach nocnych. Skoki następują w godzinach dojazdu i powrotu z pracy.

Box And Whisker Plots względem klasy



Rys. 8.1. Box and Whisker plot dla prędkości wiatru względem klasy Bike.Count.Group

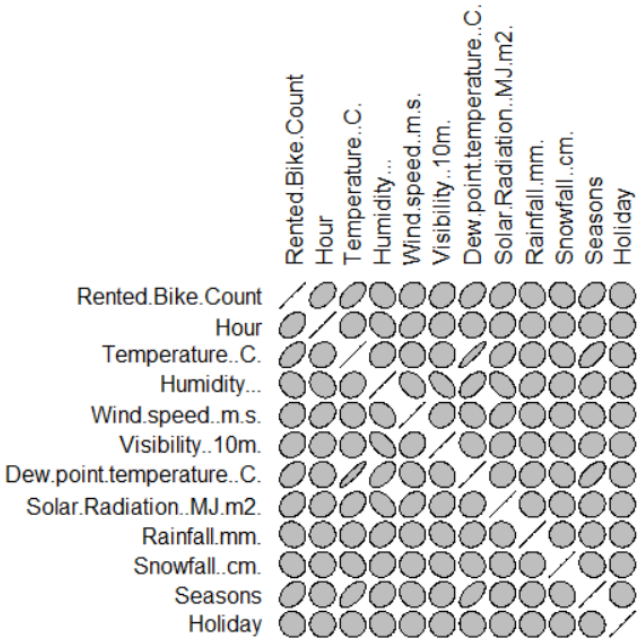
Na powyższym wykresie widzimy, że klasy nakładają się, więc prędkość wiatru nie ma dużego wpływu na ilość wypożyczonych rowerów.



Rys. 8.2. Box and Whisker plot dla promieniowania słonecznego względem klasy Bike.Count.Group

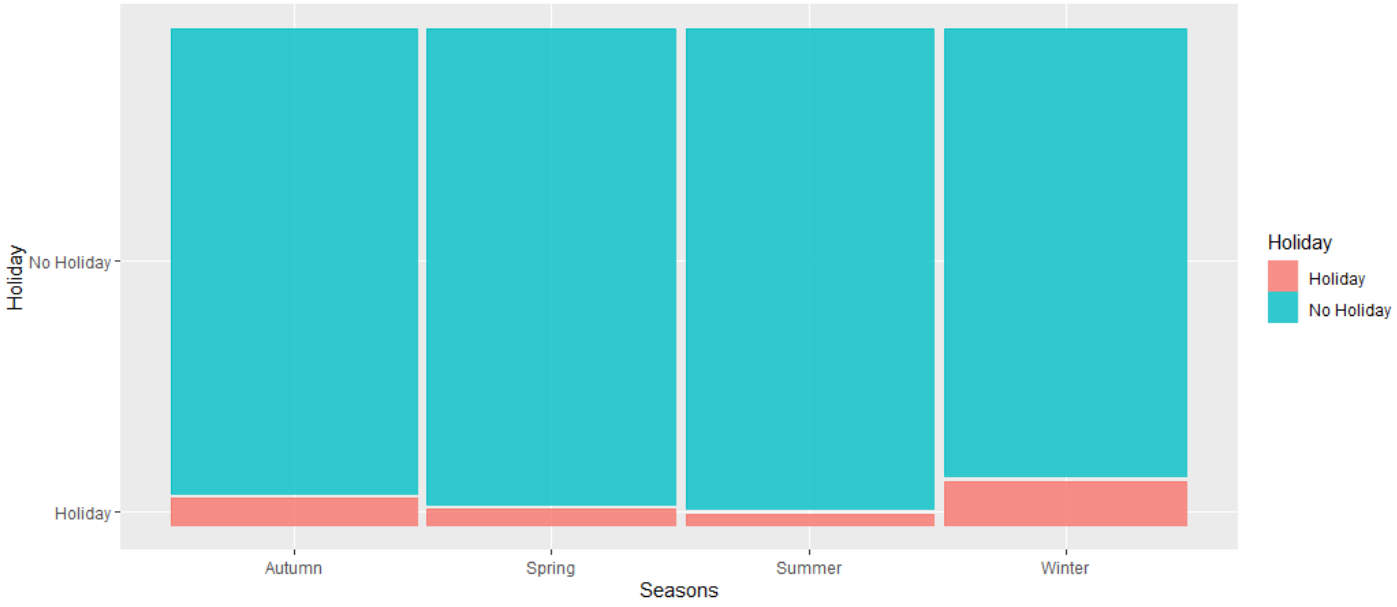
Tutaj z kolei widzimy, że nie wszystkie boxploty się nakładają, co jest logiczne, gdyż niskie promieniowanie słoneczne występuje w nocy, kiedy częstotliwość wypożyczanych rowerów drastycznie spada.

Zwizualizowaliśmy także związki między atrybutami za pomocą pakietu ellipse(wnioski, które możemy z niego wyciągnąć są analogiczne do tych z Correlation plot).



Rys. 9. Macierzy korelacji pomiędzy atrybutami z wykorzystaniem pakietu ellipse

Wykresy mozaikowe:



Rys 10.1. Wykres mozaikowy pomiędzy dniami świątecznymi a porami roku

Na powyższym wykresie widzimy, że najwięcej dni świątecznych w Seoulu występuje zimą i jesienią.

Rys. 10.2. Wykres mozaikowy pomiędzy godzinami otwarcia wypożyczalni a porami roku

Na powyższym wykresie widzimy, że wypożyczalnia była wyłączona z użytku najdłużej jesienią i wiosną. Mogło to być związane z pracami konserwacyjnymi przed i po sezonie.

Podsumowanie

W trakcie EDA wybranego zbioru danych udało nam się zauważyć sporo ciekawych związków i zależności. Dotyczyły one nie tylko wpływu różnorakich atrybutów na ilość wypożyczonych rowerów, ale także zależności między warunkami pogodowymi i zjawiskami fizycznymi. Sporo atrybutów jest ze sobą powiązanych, więc część z nich można by ze sobą połączyć bądź usunąć (głównie te dotyczące zjawisk pogodowych). Podsumowując, zgodnie z przypuszczeniami okazało się, że ładna pogoda sprzyja wypożyczaniu rowerów. Duży wpływ ma także godzina – piki w godzinach dojazdu i powrotu z pracy. Dowiedzieliśmy się, że w Seoulu nie występuje sporo opadów atmosferycznych. Co ciekawe w dni świąteczne ludność także używała rowerów miejskich, nie służą więc one tylko jako transport do pracy.