**Sprawozdanie z projektu zaliczeniowego z przedmiotu „Programowanie w R i wizualizacja danych”**

Dane, na których przeprowadzona była analiza oraz wizualizacja to plik SeulBikeData.csv przedstawiający obserwacje roczne od 1 grudnia 2017 do 30 listopada 2018 roku wynajętych rowerów w zależności od dnia oraz zjawisk pogodowych.

Dane zawarte w pliku przedstawiają informację o:

- dniu pomiarów,

- ilości wypożyczonych rowerów,

- czasie pracy wypożyczalni,

- temperatury,

- zachmurzenia,

- prędkości wiatru,

- widoczności,

- temperatury punktu rosy,

- promieniowaniu słonecznym,

- opadach deszczu,

- opadach śniegu,

- porze roku,

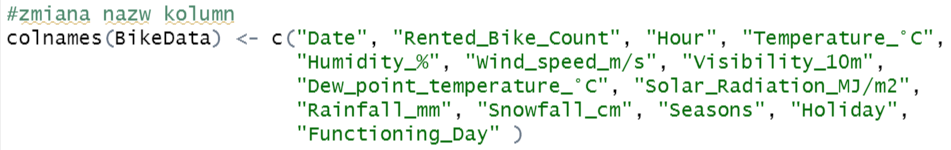
- czy był to okres wakacji,

- czy wypożyczalnia była w danym dniu otwarta.

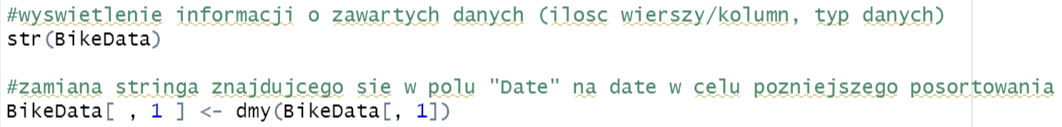
Do przeprowadzenia analizy danych oraz ich sformatowania, użyte zostały pakiety *„dplyr”, „lubridate”* oraz *„tidyverse”*. Za pomocą funkcji *read.csv* dane z pliku zostały przypisane do zmiennej *BikeData.*



Następnie poprawione zostały nazwy kolumn, tak, aby nie zawierały w sobie spacji.



Funkcją *str()* sprawdzony został typ danych. Można zauważyć, że informacja o dniu została przedstawiona jako tekst, a nie format daty, dlatego dane z kolumny *„Date”* zostały przekształcone za pomocą funkcji *dmy()* z pakietu *lubridate* na typ data.



Aby lepiej przyjrzeć się danym, zastosowane zostały takie funkcje jak:

- *head()* – wyświetla 6 pierwszych wierszy danych,

- *arrange()* – sortuje dane; dla tego projektu najpierw według daty pomiaru, a następnie według

ilości wypożyczonych rowerów,

- *table()* – zlicza ilość występowania rodzajów danych,

- *is.na()* – sprawdza, czy nie występuje brak wartości dla danego pomiaru,

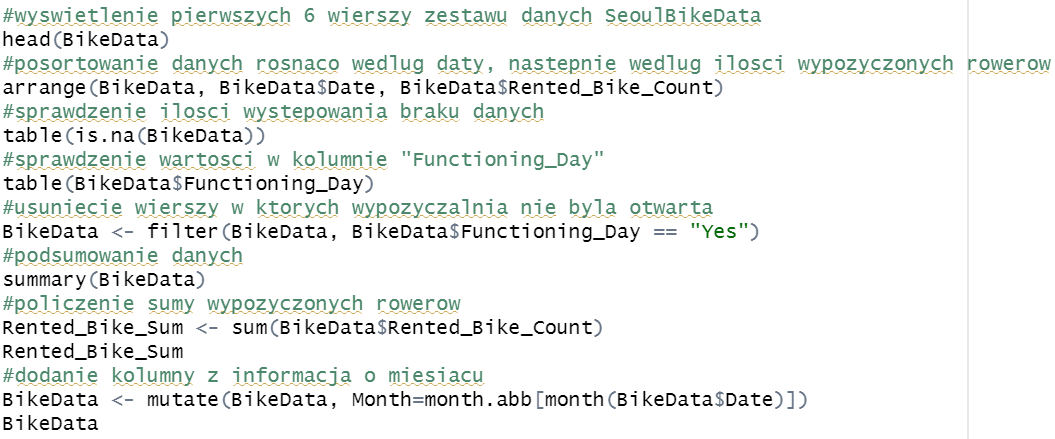
- *filter()* – wybiera z pliku wiersze spełniające warunek,

- *summary()* – podsumowanie danych; wyświetla informację dla każdej kolumny o jej minimalnej

i maksymalnej wartości, średniej oraz medianie, bądź typie danych i ilości wierszy

- *sum()* – zlicza wartość całej kolumny danych,

- *mutate()* – pozwala utworzyć nową kolumnę danych w pliku



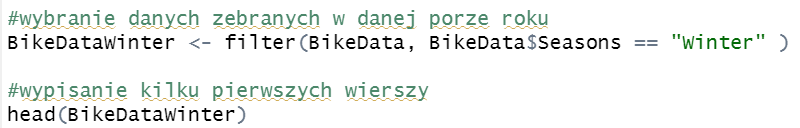
Powyższe funkcje pokazały, że w pliku nie występują braki danych. Znajdują się natomiast informacje o wypożyczonych rowerach w dniach, gdy wypożyczalnia była zamknięta. Dane te zostały usunięte z pliku tak, aby dalsza analiza przeprowadzona została na pomiarach przeprowadzonych w dniach, w których wypożyczalnia funkcjonowała.

Ponadto wyznaczona została całkowita ilość sprzedanych rowerów wynosząca ponad 6mln. Do danych została dołożona kolumna z informacją o miesiącu pomiaru, w celu przeprowadzenia wizualizacji w późniejszym etapie projektu.

Tak przekształcone dane zostały zapisane do pliku funkcją *write.csv().*



Ze względu na dokładniejsze przedstawienie poszczególnych danych oraz ich lepszą wizualizację, zostały one podzielone ze względu na pory roku. Dla tak podzielonych obserwacji została przeprowadzona analiza danych za pomocą funkcji z pakietu *tidyverse* oraz *dplyr*.



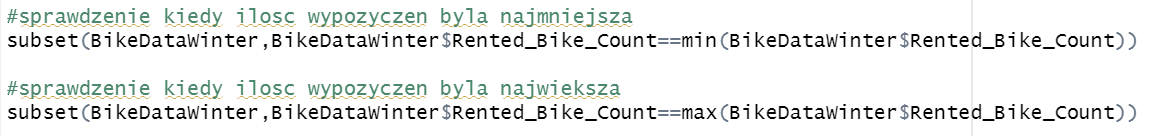
Najpierw zostały zliczone wiersze dla każdej z pór roku. Najwięcej zebranych pomiarów zawierają dane z lata (2208 wierszy), dalej w kolejności znajdują się pomiary z zimy oraz wiosny (2160 pomiarów), najmniej - bo tylko 1937 wierszy – dane z jesieni.



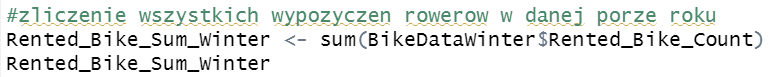
Następnie sprawdzona została sytuacja, gdy wypożyczalnia mimo bycia otwartą w dany dzień nie wypożyczyła żadnego roweru. Funkcja *filter()* nie znalazła takich obserwacji dla żadnej z pory roku.



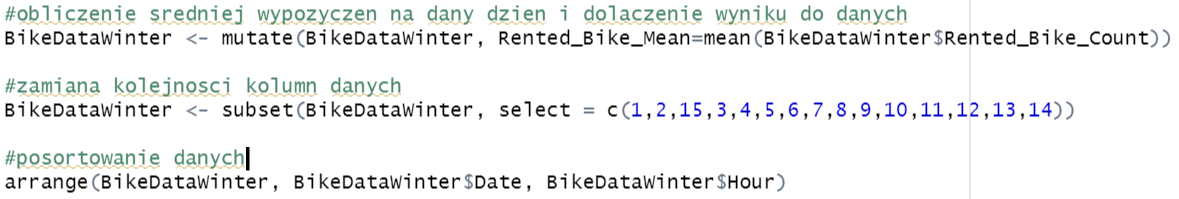
Kolejnym krokiem było wyznaczenie wiersza z najmniejszą oraz największą liczbą wypożyczonych rowerów.



Nasuwa się także pomysł policzenia, w której porze roku zostało wypożyczonych najwięcej rowerów. Na pierwszym miejscu znajduje się lato, czyli okres wakacji, z ilością ponad 2mln wypożyczonych rowerów. Następnie dość zbliżone do siebie w wynikach wiosna oraz jesień (wiosna 1,6mln; jesień 1,8mln). Na samym końcu, zgodnie z oczekiwaniami, plasuje się zima z wynikiem 0,5mln.

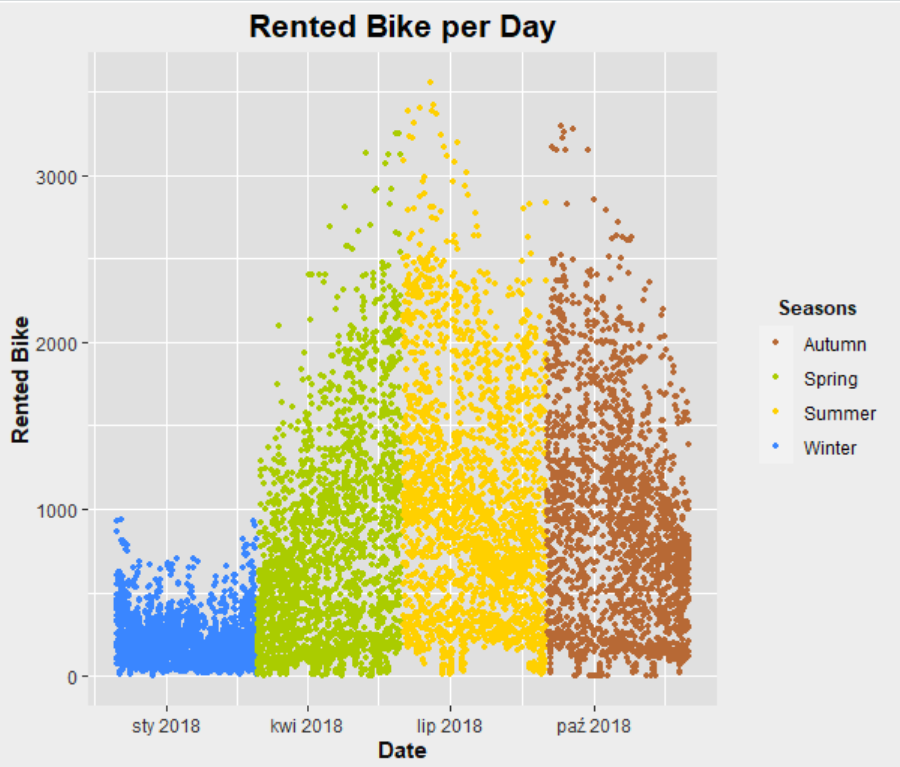


Kolejnymi operacjami przeprowadzonymi na danych jest wyliczenie średniej ilości wypożyczonych rowerów w danym dniu oraz przypisanie wyniku do danych, zamiana kolejności kolumn oraz posortowanie danych według daty i godzin otwarcia wypożyczalni.



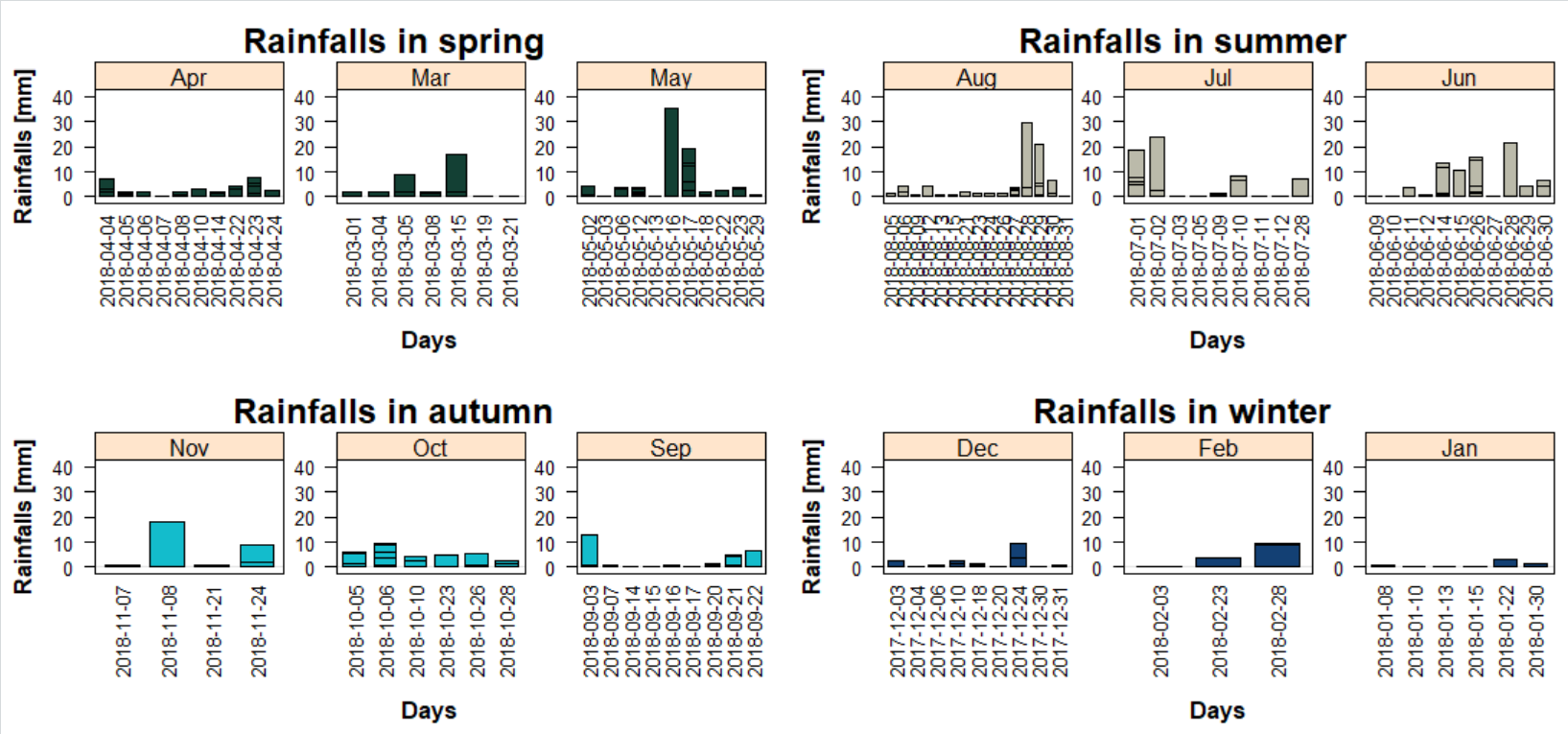
W celu lepszego przedstawienia danych, zostało przeprowadzonych pięć wizualizacji z pięciu różnych pakietów:

1. Pakiet *ggplot2* – zależność wypożyczonych rowerów w danej porze roku od dnia.



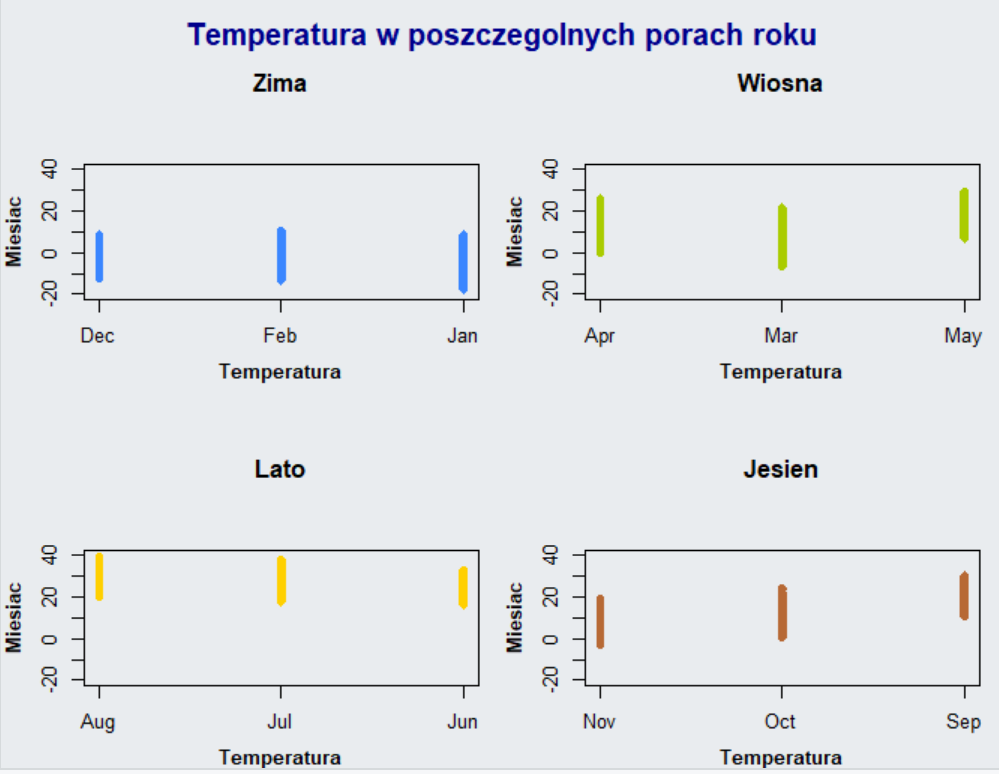
Zgodnie z wcześniejszą analizą danych – najwięcej rowerów zostało wypożyczonych w lecie. Co więcej, również w tej porze roku znajduje się największa ilość wypożyczonych rowerów w danym dniu. Wiosna oraz jesień są do siebie bardzo zbliżone pod względem wyników – to praktycznie lustrzane odbicie. Według informacji na wykresie, można też zaobserwować okres wakacji – zdecydowany wzrost wypożyczonych rowerów następuje w okolicach kwietnia, trwa przez całe lato i kończy się w okolicach września.

1. Pakiet *lattice* – ilość opadów w zależności od dnia, z wyszczególnieniem pór roku.



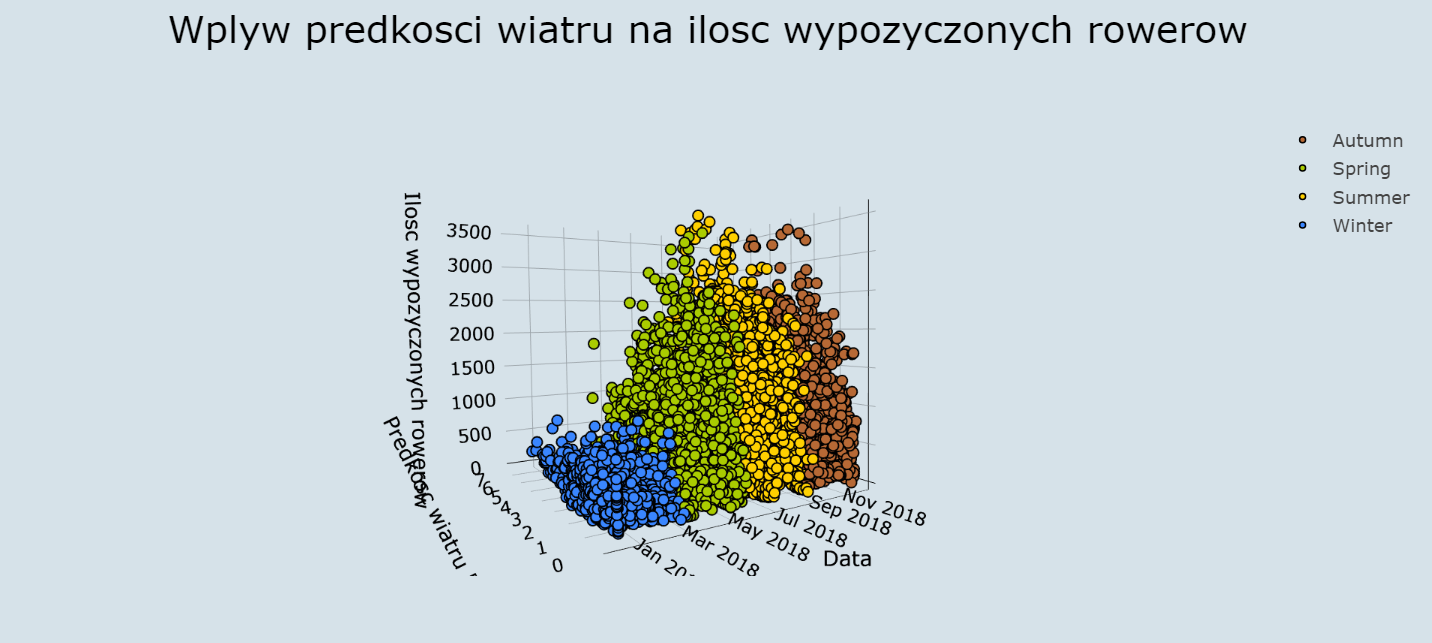
Dla większej czytelności wykresu, jako dane na osi X zostały wybrane te dni, dla których występowały jakiekolwiek opady deszczu – brak dni z opadami równymi 0mm. Najwięcej opadów nastąpiło w lecie, najmniej – w zimie. Miesiącem z największą ilością dni deszczowych był sierpień. Najwięcej opadów w ciągu dnia spadło 16 maja 2018 roku, czyli na wiosnę.

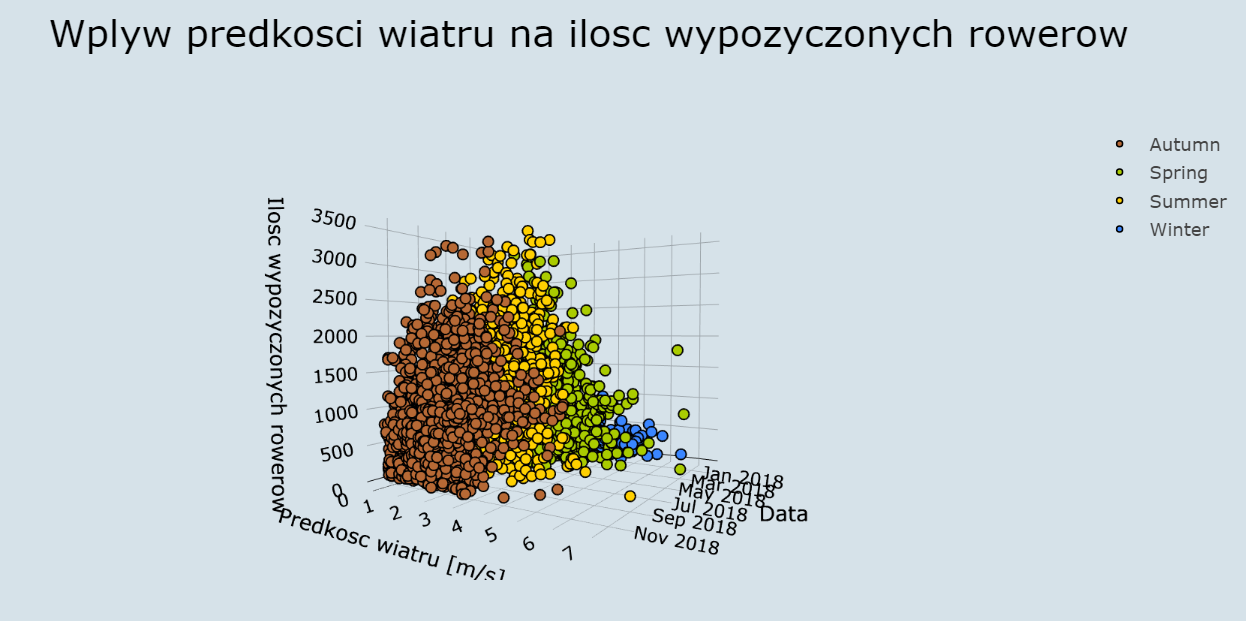
1. Pakiet *graphics* – zależność temperatury od pory roku



Zgodnie z oczekiwaniami – najniższa temperatura występowała zimą – od prawie -20 stopni do maksymalnie 10 stopni. Najwyższa natomiast latem – od 20 stopni do aż 40 stopni. Najmniejsze wahania temperatury nastąpiły latem, największe – wiosną. W zimie oraz w lecie rozpiętość temperatury utrzymywała się praktycznie na stałym poziomie w każdym miesiącu (dla zimy od około -15 stopni do 10 stopni; dla lata od około 20 stopni do 40 stopni). Dla wiosny można zaobserwować wzrost temperatury wraz z kolejnymi miesiącami, dla jesieni – spadek.

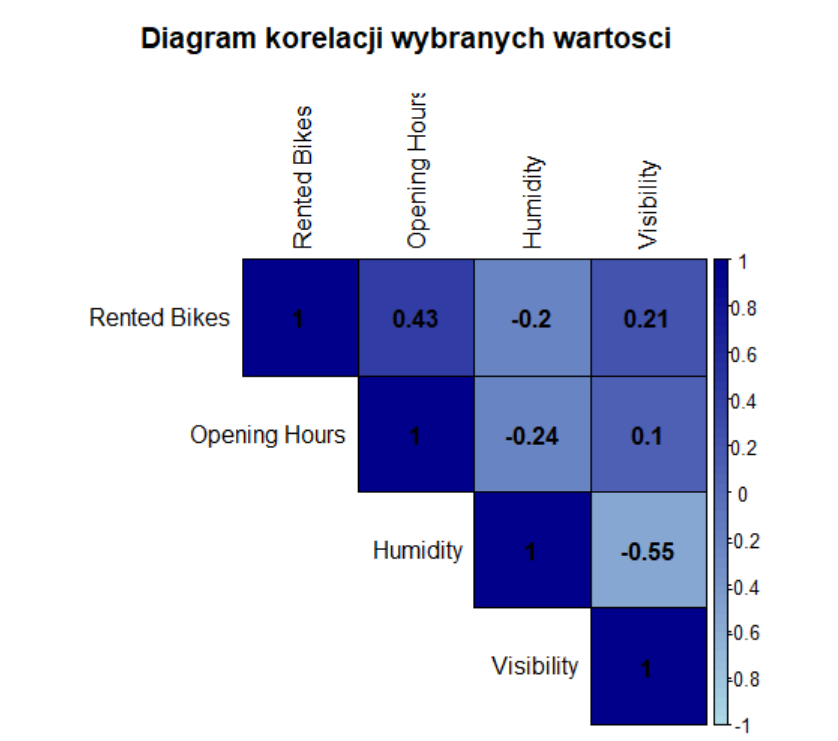
1. Pakiet *plotly* - zależność ilości wypożyczonych rowerów od prędkości wiatru w danym dniu





Z wykresu dowiadujemy się, że najwięcej rowerów zostało wypożyczonych w lecie, najmniej – w zimie. Ponadto obserwujemy zależność prędkości wiatru od ilości wypożyczonych rowerów. Najwięcej rowerów zostało wypożyczonych przy maksymalnej prędkości wiatru 3m/s. Następnie zagęszczenie punktów na wykresie drastycznie spada ku podłożu. Wyjątkiem są dwa dni, gdy wiatr wiał z prędkością 6m/s, a ilość wypożyczonych rowerów sięgnęła 1000 do 2000 sztuk.

1. Pakiet *corrplot* – diagram korelacji danych liczbowych z pliku SeulBikeData.csv



Ostatnią wizualizacją danych jest diagram korelacji dla wartości liczbowych zawartych w pliku. Widać, że największy wpływ na ilość wypożyczonych rowerów ma czas otwarcia wypożyczalni (współczynnik korelacji : 0,43). Praktycznie brak wpływu na siebie mają ilość wypożyczonych rowerów oraz zachmurzenie i godziny otwarcia oraz widoczność. Ze zjawisk pogodowych, największa zależność występuje pomiędzy widocznością a zachmurzeniem. Wraz ze zmniejszeniem się zachmurzenia wzrasta widoczność.