

Лабораторная работа 4. Графический анализ данных

Цель занятия — научиться выполнять анализ данных в R с помощью операций визуализации из базового пакета и ggplot2.

Вспомогательный материал: Лекция 5.

Общие указания:

1. В качестве отчета о выполнении практического занятия студент представляет преподавателю программный код (*lr4.R*).
2. Для импорта, экспорта и обработки данных нельзя использовать возможности интерфейса RStudio — можно только писать программный код.
3. Задания должны быть выполнены наиболее оптимальным образом (автоматизация, минимизация строк кода, универсальность и др.).
4. Программный код должен быть оформлен в соответствии с Google's R Style Guide.
5. Смысловые блоки программного кода необходимо сопровождать комментариями.

Задание 1

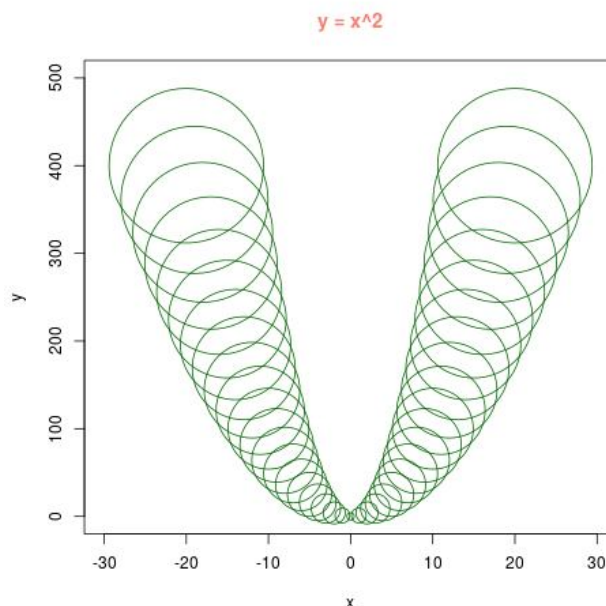
—> *visualisation.R*

<— *lr4_1.png*

Откройте скрипт *visualisation.R*. Ознакомьтесь с кодом.

Создайте скрипт *lr4.R* для выполнения задания.

Создайте с помощью функции **plot()** идентичный график:



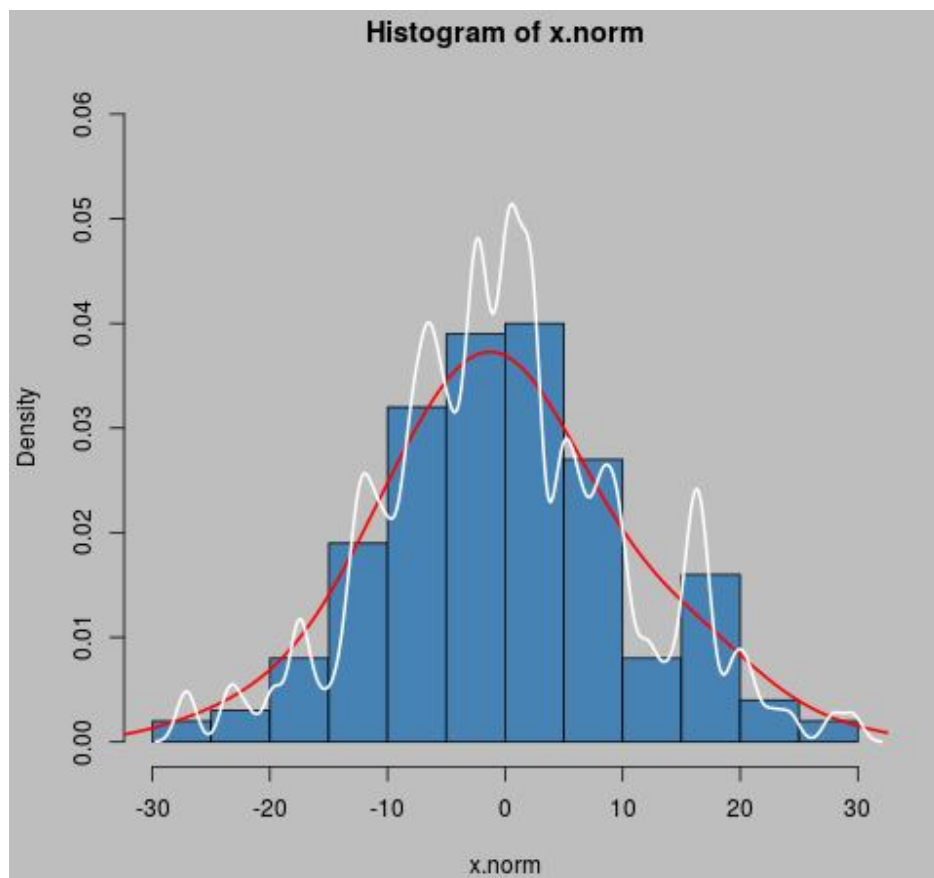
Сохраните его в папке Plots под именем *lr4_1.png*

Задание 2

`<- lr4_2.png`

Сформируйте нормально распределенную совокупность `x.norm` из 200 наблюдений со средним значением 0 и стандартным отклонением 10.

С помощью функций `png()`, `par()`, `hist()`, `lines()`, `density()`, `dev.off()` создайте график с такими же параметрами, как на рисунке ниже, и сохраните его в папке Plots под именем `lr4_2.png`. При построении графика задайте разбиение гистограммы на 15 частей. Проанализируйте, почему гистограмма разделена на иное количество частей (если это так).



Задание 3

Загрузите dataset `mpg` из пакета `ggplot2`. Ознакомьтесь с переменными.

1. Постройте диаграмму рассеяния `hwy` от `cyl`.

Постройте такую же диаграмму с параметром `position = "jitter"`, проанализируйте разницу.

2. Почему в графике, построенном с помощью следующего кода, точки не синие? Что нужно исправить?

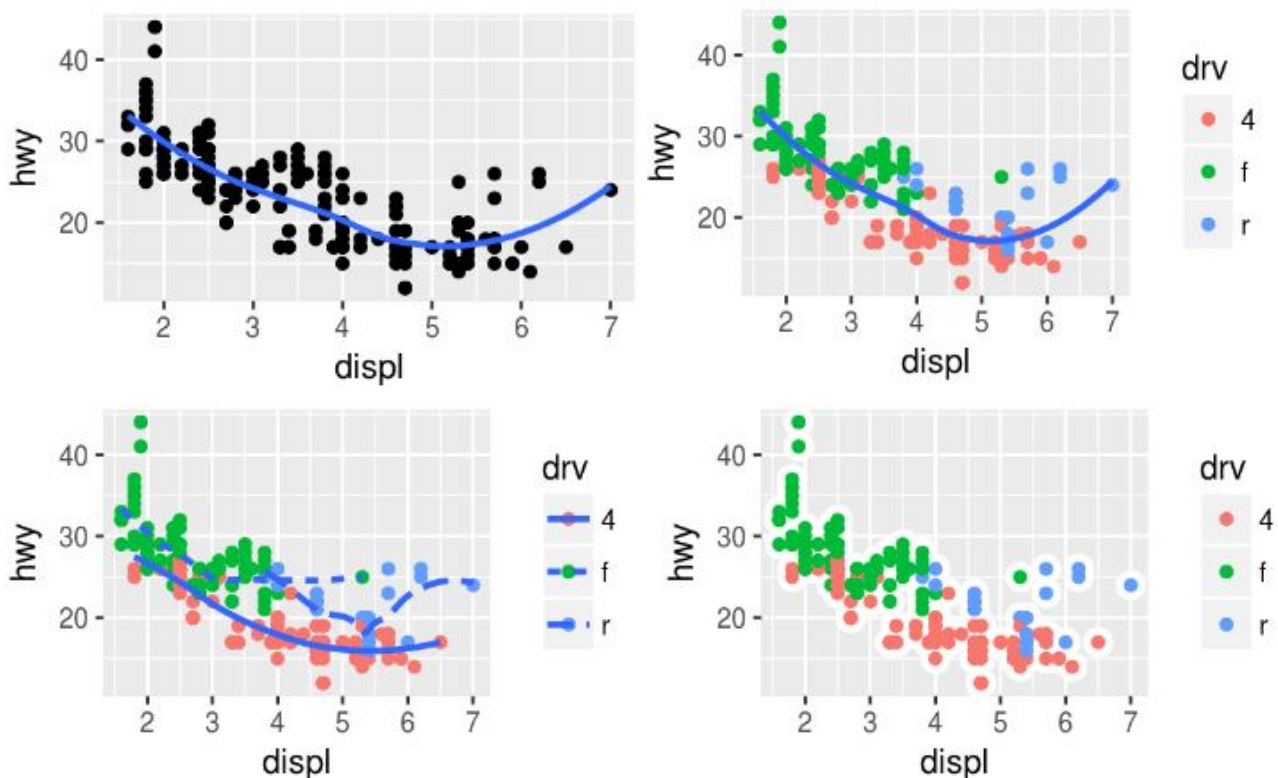
```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy, color = "blue"))
```

- Отобразите через параметры `color`, `size`, и `shape` поочередно количественные и качественные переменные. Сравните, как `aesthetics` ведет себя в каждом случае.
- Что будет, если отобразить через несколько параметров `aesthetics` одну и ту же переменную?
- Для чего предназначена `stroke` aesthetic?
- Что будет, если отобразить через `aesthetic` не переменную, а, например, `aes(colour = displ < 5)`?
- Что произойдет, если построить `facet` по количественной переменной?
- Какой параметр задает `.` (точка) для следующих графиков?

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy)) +  
  facet_grid(drv ~ .)
```

```
ggplot(data = mpg) +  
  geom_point(mapping = aes(x = displ, y = hwy)) +  
  facet_grid(. ~ cyl)
```

- Для чего используется параметр `se` в `geom_smooth()`?
- Напишите код для создания следующих графиков:



11. Для чего предназначен `geom_col()`, чем он отличается от `geom_bar()`?

12. Как можно улучшить следующий график?

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +  
  geom_point()
```

13. Преобразуйте stacked bar chart в круговую диаграмму с помощью `coord_polar()`.

14. В чем разница между `coord_quickmap()` и `coord_map()`?

15. Какие выводы можно сделать из следующего графика о взаимоотношении потребления топлива на трассе и в городе? Почему параметр `coord_fixed()` важен? Для чего предназначен `geom_abline()`?

```
ggplot(data = mpg, mapping = aes(x = cty, y = hwy)) +  
  geom_point() +  
  geom_abline() +  
  coord_fixed()
```

Литература:

Grolemund, G. *R for Data Science [Electronic resource]* / Garrett Grolemund, Hadley Wickham. – 2016. – Mode of access: <http://r4ds.had.co.nz/index.html>. – Date of access: 01.09.2016.