

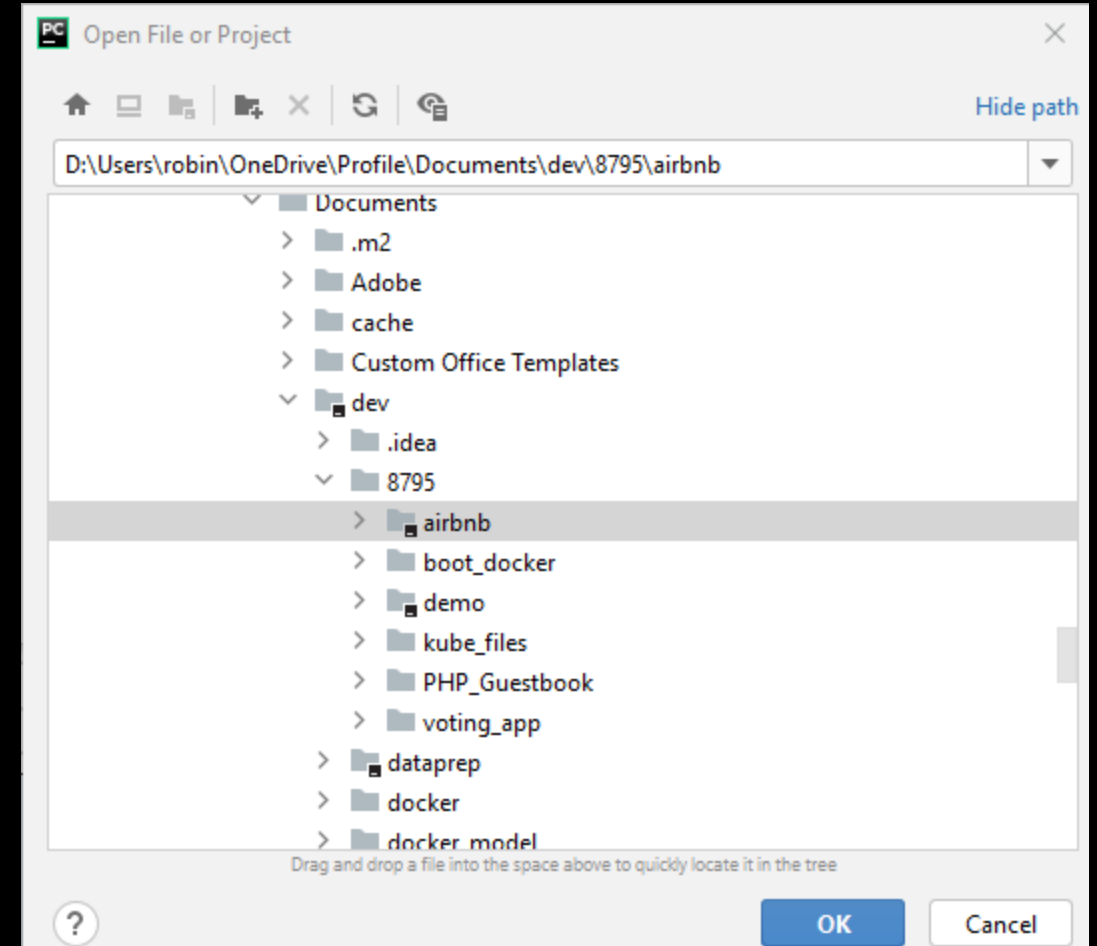
Run PySpark Airbnb locally

Steps to run Airbnbnb

- Install PySpark interpreter locally
 - See other presentation
- Run in PyCharm
 - Play button
 - Python interpreter
- Run in command line
- Run in Kubernetes (Docker for Desktop)

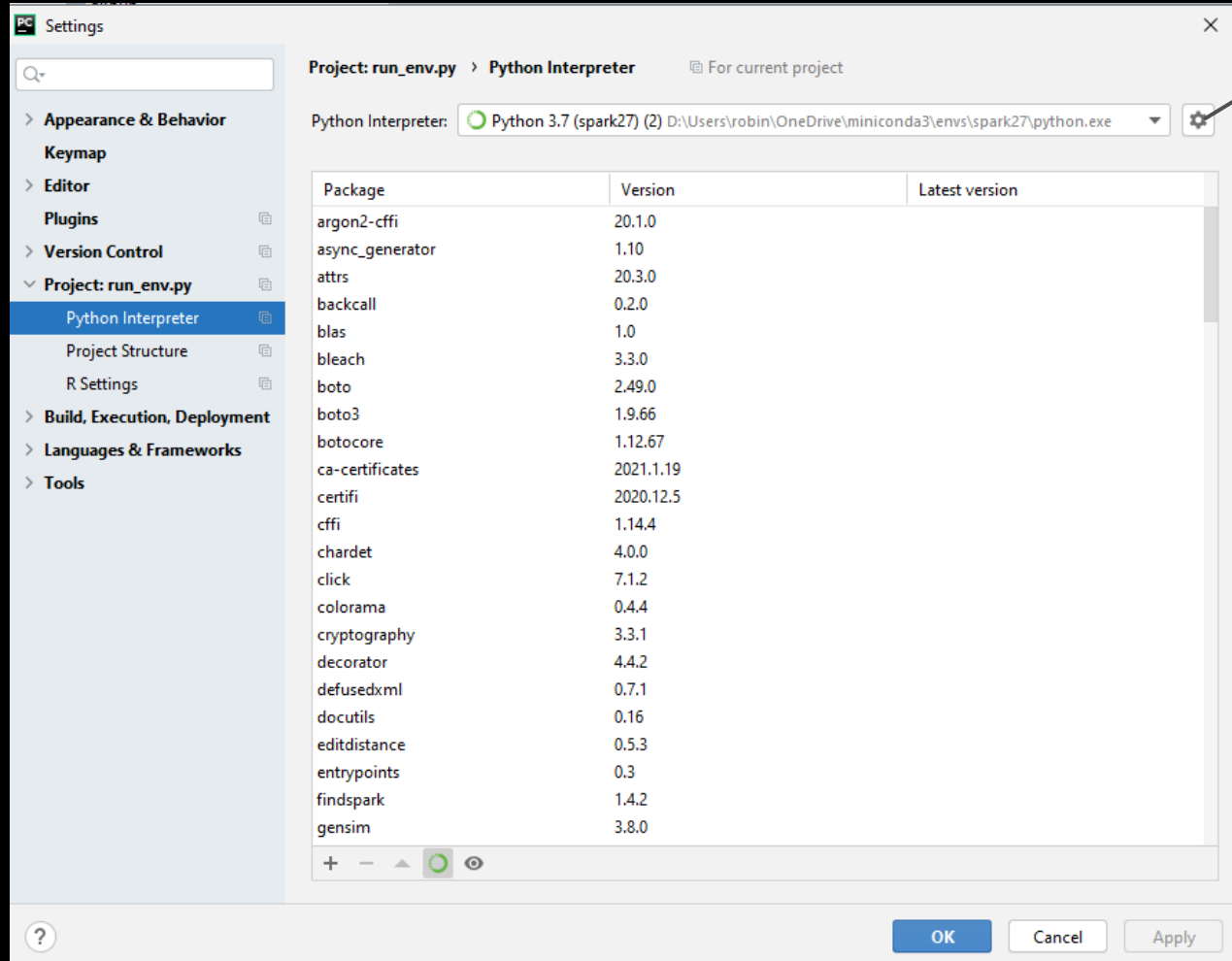
Run in PyCharm

- In PyCharm
 - Open project by selecting its directory
 - Topmost airbnb directory



Ensure project has PySpark interpreter

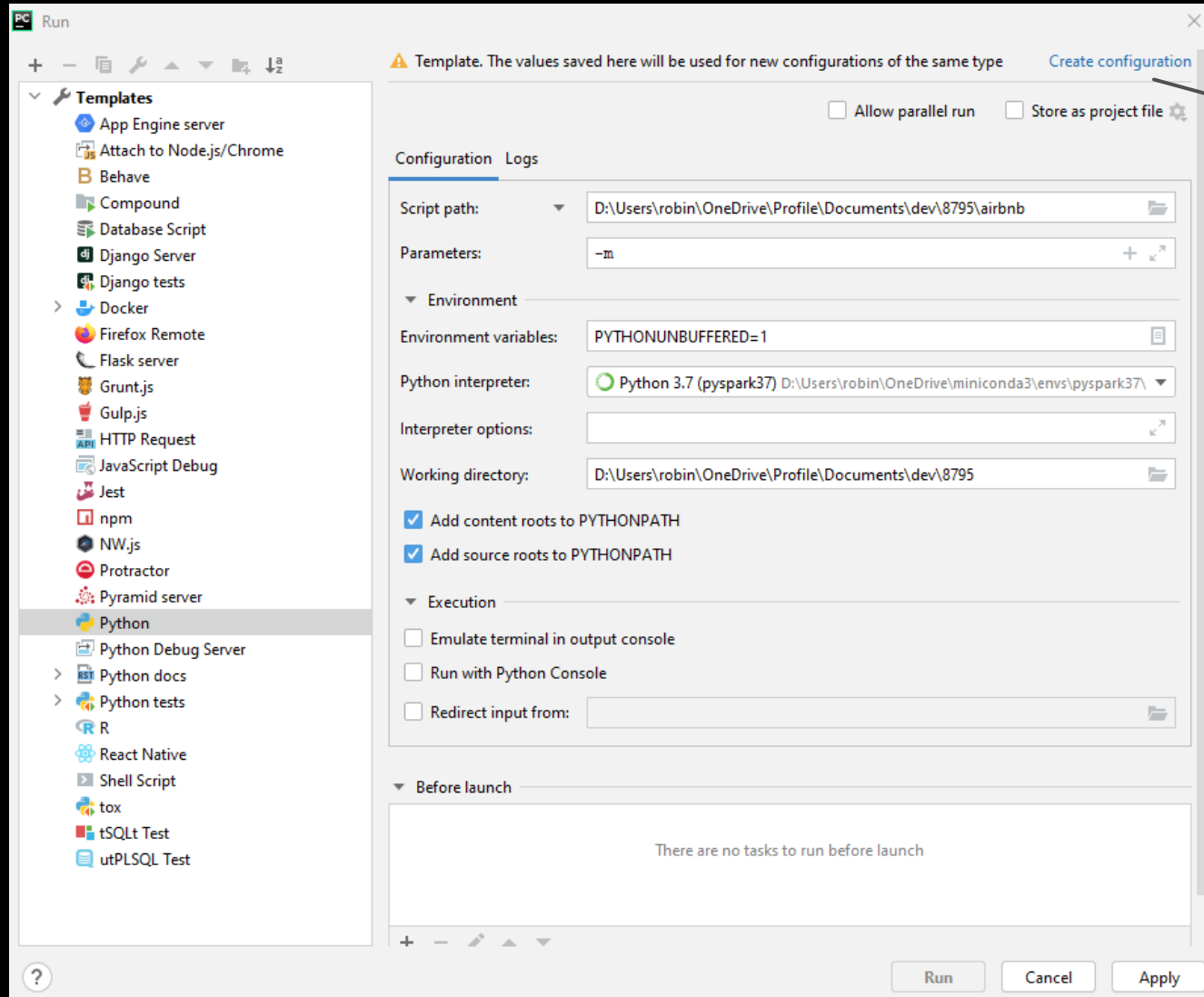
See other presentation



Gear
Add Interpreter

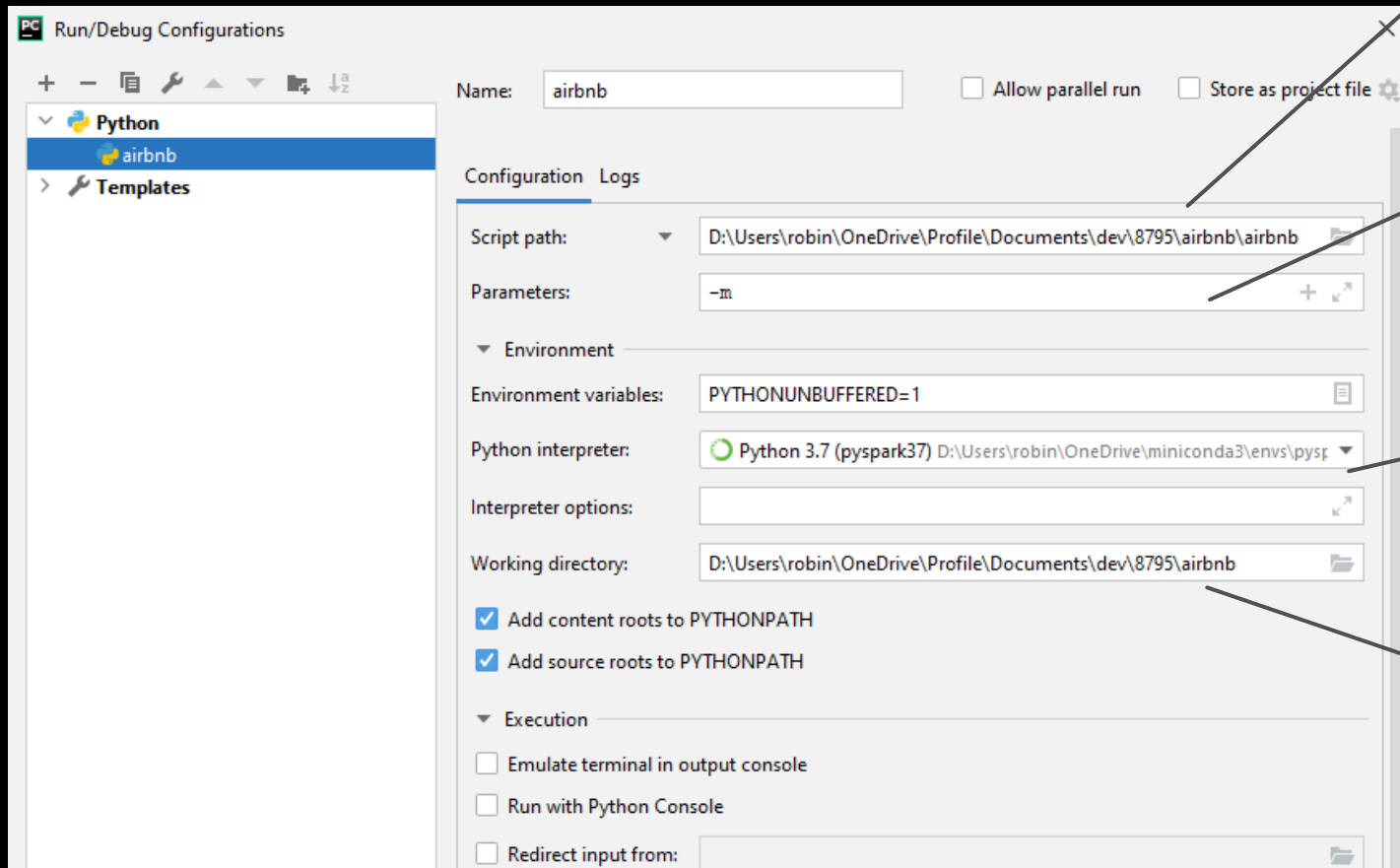
PyCharm Run Configuration

Run | Edit Configurations | Templates | Python



Create configuration

Configuration arguments



Local airbnb/airbnb
src directory

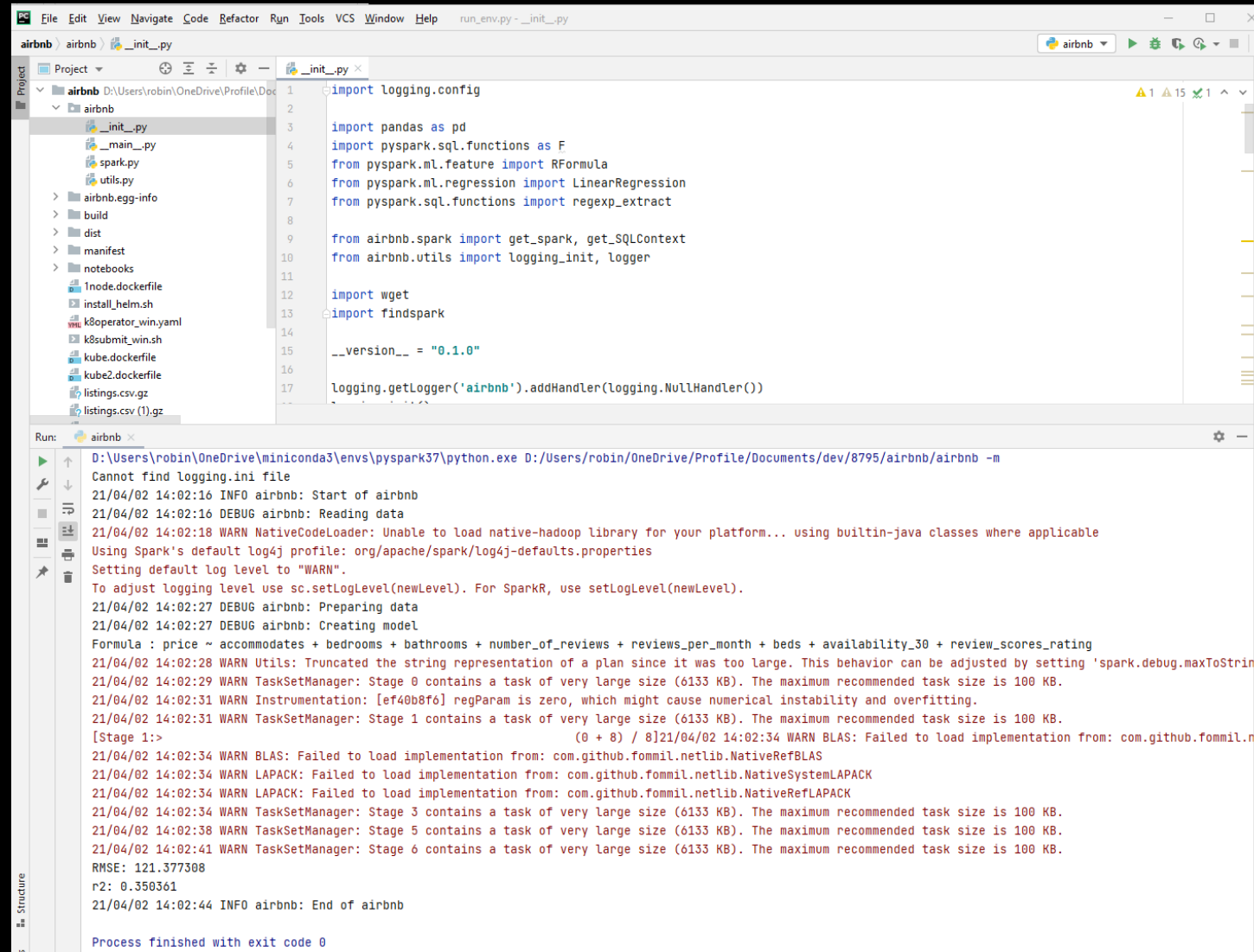
-m

Already set to
interpreter

Local airbnb
directory

Now, run projects in PySpark interpreter

Play button upper right



The screenshot shows an IDE window with the following components:

- Project Explorer (Left):** Displays the project structure for 'airbnb'. The file `_init_.py` is selected.
- Code Editor (Center):** Shows the contents of `_init_.py`, which includes imports for `logging.config`, `pandas`, `pyspark.sql.functions`, `RFormula`, `LinearRegression`, `regexp_extract`, `get_spark`, `get_SQLContext`, `wget`, and `findspark`. It also sets `__version__ = "0.1.0"` and configures logging.
- Run Console (Bottom):** Shows the output of running the project. The command executed is `D:\Users\robin\OneDrive\Profile\Documents\dev\8795\airbnb\airbnb -m`. The output includes logs for starting the project, reading data, creating a model, and training the model. The final output shows the RMSE (121.377308) and R2 (0.350361) values.

```
import logging.config
import pandas as pd
import pyspark.sql.functions as F
from pyspark.ml.feature import RFormula
from pyspark.ml.regression import LinearRegression
from pyspark.sql.functions import regexp_extract

from airbnb.spark import get_spark, get_SQLContext
from airbnb.utils import logging_init, logger

import wget
import findspark

__version__ = "0.1.0"

logging.getLogger('airbnb').addHandler(logging.NullHandler())
```

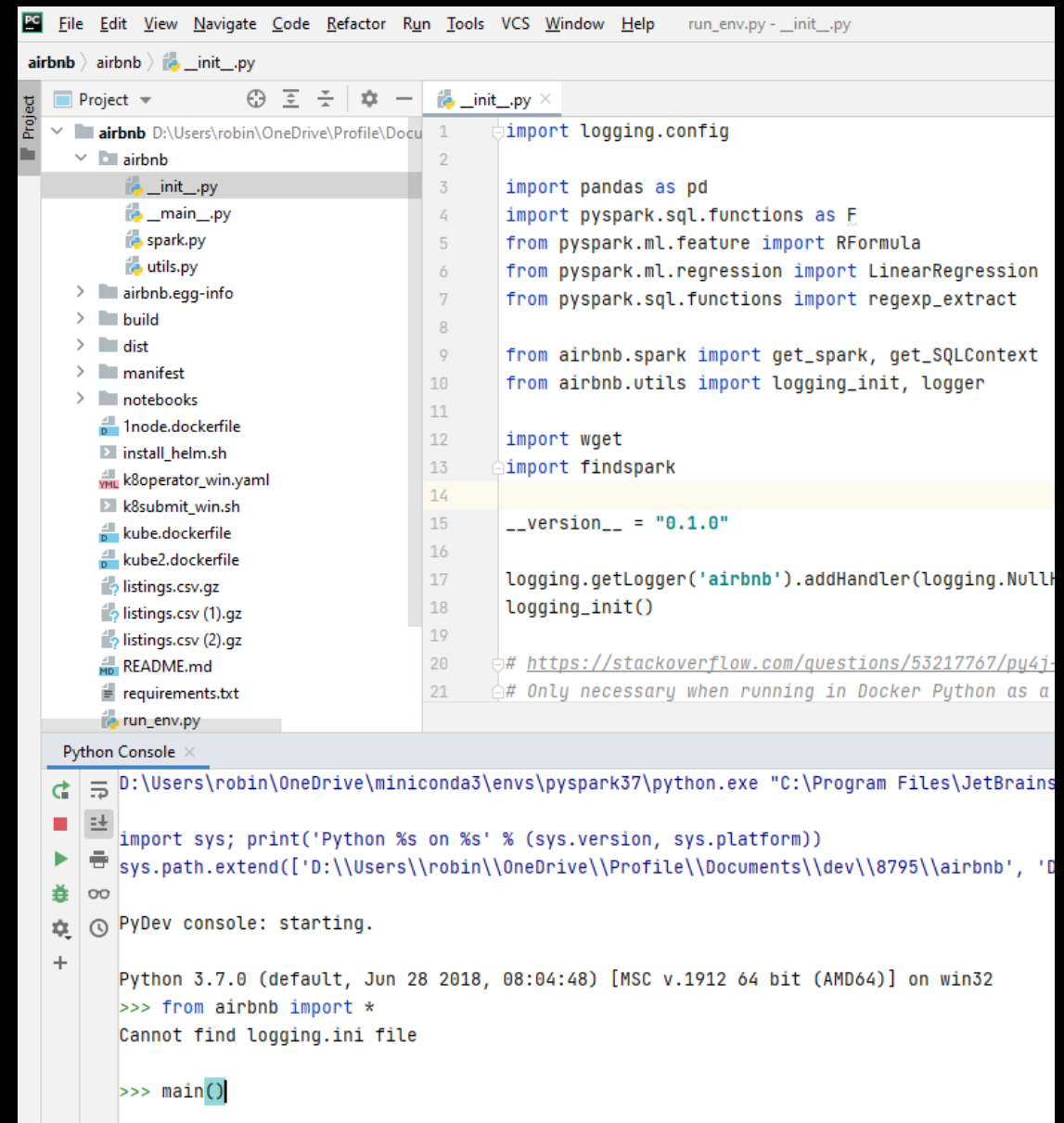
Run: airbnb

```
D:\Users\robin\OneDrive\miniconda3\envs\pyspark37\python.exe D:/Users/robin/OneDrive/Profile/Documents/dev/8795/airbnb/airbnb -m
Cannot find logging.ini file
21/04/02 14:02:16 INFO airbnb: Start of airbnb
21/04/02 14:02:16 DEBUG airbnb: Reading data
21/04/02 14:02:18 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
Setting default log level to "WARN".
To adjust logging level use sc.setLogLevel(newLevel). For SparkR, use setLogLevel(newLevel).
21/04/02 14:02:27 DEBUG airbnb: Preparing data
21/04/02 14:02:27 DEBUG airbnb: Creating model
Formula : price ~ accommodates + bedrooms + bathrooms + number_of_reviews + reviews_per_month + beds + availability_30 + review_scores_rating
21/04/02 14:02:28 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringLength'.
21/04/02 14:02:29 WARN TaskSetManager: Stage 0 contains a task of very large size (6133 KB). The maximum recommended task size is 100 KB.
21/04/02 14:02:31 WARN Instrumentation: [ef40b8f6] regParam is zero, which might cause numerical instability and overfitting.
21/04/02 14:02:31 WARN TaskSetManager: Stage 1 contains a task of very large size (6133 KB). The maximum recommended task size is 100 KB.
[Stage 1:]
(0 + 8) / 8]21/04/02 14:02:34 WARN BLAS: Failed to load implementation from: com.github.fommil.netlib.NativeRefBLAS
21/04/02 14:02:34 WARN LAPACK: Failed to load implementation from: com.github.fommil.netlib.NativeSystemLAPACK
21/04/02 14:02:34 WARN LAPACK: Failed to load implementation from: com.github.fommil.netlib.NativeRefLAPACK
21/04/02 14:02:34 WARN TaskSetManager: Stage 3 contains a task of very large size (6133 KB). The maximum recommended task size is 100 KB.
21/04/02 14:02:38 WARN TaskSetManager: Stage 5 contains a task of very large size (6133 KB). The maximum recommended task size is 100 KB.
21/04/02 14:02:41 WARN TaskSetManager: Stage 6 contains a task of very large size (6133 KB). The maximum recommended task size is 100 KB.
RMSE: 121.377308
r2: 0.350361
21/04/02 14:02:44 INFO airbnb: End of airbnb

Process finished with exit code 0
```

Run in PyCharm interpreter

- Open python console
 - `from airbnb import *`
 - `main()`



Run in command line

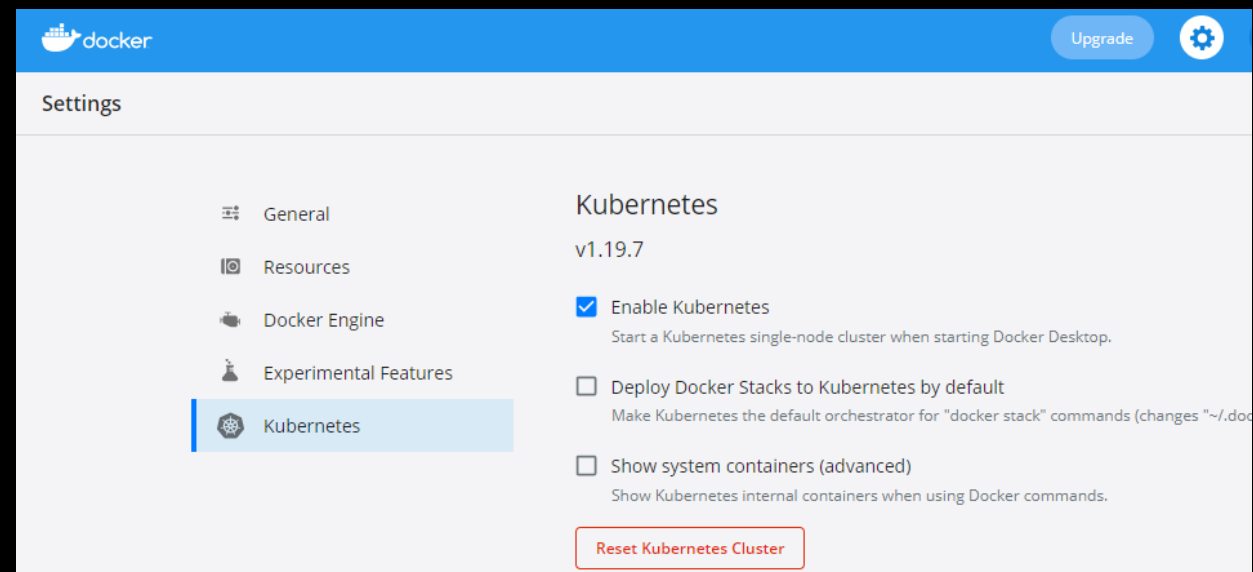
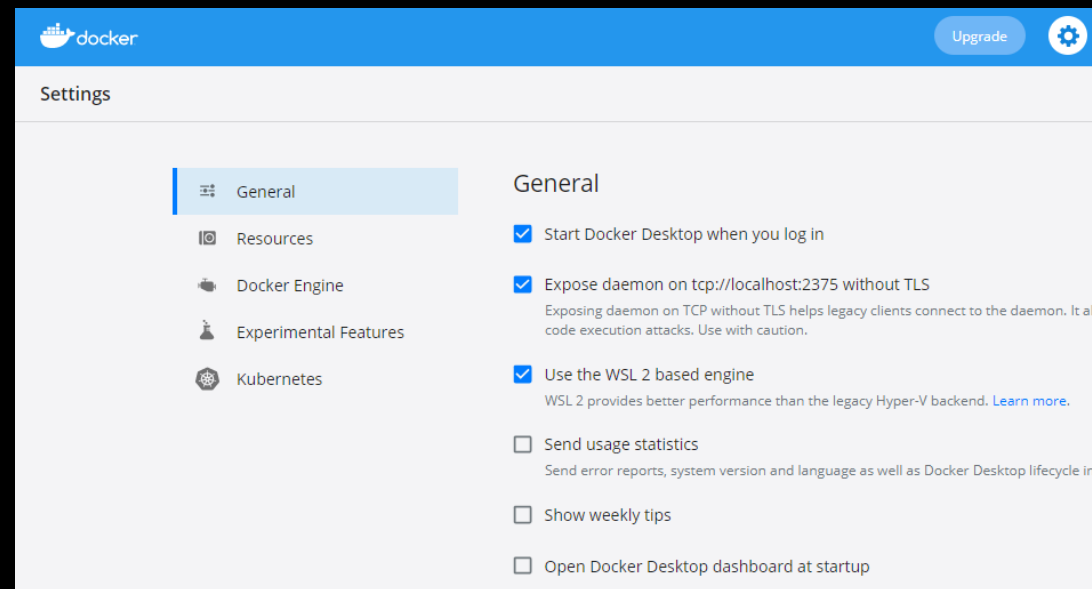
- Open shell
 - Windows command, Windows WSL 2, ...
- Change directory to airbnb
- Run in python
 - `conda activate pyspark37`
 - If using conda, ensure environment is active
 - `python -m airbnb`
- In case of failure
 - Check environment variables for paths
 - `SPARK_HOME`
 - Python interpreter

Run in command line

```
(pyspark37) D:\Users\robin\OneDrive\Profile\Documents\dev\8795\airbnb>python -m airbnb
Cannot find logging.ini file
21/04/02 14:31:35 INFO airbnb: Start of airbnb
21/04/02 14:31:35 DEBUG airbnb: Reading data
100% [.....] 9032702 / 9032702
N NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-jav
e
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
```

Run in Kubernetes (Docker for Desktop)

- Assuming Docker for Desktop is installed
 - and Kubernetes is enabled



Run in Kubernetes (Docker for Desktop)

- Directory where airbnb is copied
 - In WSL2 (unix)
 - `cd /mnt/c/airbnb`
- In one shell (Windows command or WSL2 shell)
 - `docker build -f lnode.dockerfile -t wrobinson/airbnb:1.0 .`
 - `./k8submit_win.sh`
- In another shell (or after the above completes)
 - `kubectrl logs -f -l app=airbnb --tail=-1`

Run in Kubernetes (Docker for Desktop)

```
wnr@Coolidge:airbnb$ ./k8submit_win.sh
/opt/spark/bin/spark-submit --master k8s://https://localhost:6443 --deploy-m
memory 3g --conf spark.executor.instances=2 --conf spark.dynamicAllocation.enabled
onf spark.kubernetes.driver.limit.cores=1500m --conf spark.kubernetes.executor
es.container.image=wrobinson/airbnb:1.0 --conf spark.kubernetes.driver.volum
/work-dir/shared --conf spark.kubernetes.driver.volumes.hostPath.airbnb-vol.
obin/OneDrive/Profile/Documents/dev/8795/airbnb --conf spark.kubernetes.driv
e --conf spark.kubernetes.executor.volumes.hostPath.airbnb-vol.mount.path=/c
etes.executor.volumes.hostPath.airbnb-vol.options.path=/run/desktop/mnt/host
v/8795/airbnb --conf spark.kubernetes.executor.volumes.hostPath.airbnb-vol.r
ernetes.executor.label.app=airbnb --conf spark.kubernetes.driver.label.app=a
ARK_APP_MODULE=airbnb --conf spark.kubernetes.driverEnv.PYTHONPATH=/opt/spar
rk.kubernetes.executorEnv.PYTHONPATH=/opt/spark/work-dir/shared/dist/airbnb.
--master=k8s://https://localhost:6443 --py_files=shared/dist/airbnb.zip
```

```
21/04/02 14:39:39 WARN Utils: Your hostname, Coolidge resolves to a loopback
stead (on interface eth0)
21/04/02 14:39:39 WARN Utils: Set SPARK_LOCAL_IP if you need to bind to anot
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file
.1.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spa
WARNING: Use --illegal-access=warn to enable warnings of further illegal ref
WARNING: All illegal access operations will be denied in a future release
```

```
wnr@Coolidge:8795$ kubectl logs -f -l app=airbnb --tail=-1
++ id -u
+ myuid=185
++ id -g
+ mygid=1000
+ set +e
++ getent passwd 185
+ uidentry=spark:x:185:1000::/home/spark:/bin/bash
+ set -e
+ '[' -z spark:x:185:1000::/home/spark:/bin/bash ']'
+ SPARK_CLASSPATH=':/opt/spark/jars/*'
+ env
++ id -u
+ myuid=185
++ id -g
+ mygid=1000
+ set +e
++ getent passwd 185
+ uidentry=spark:x:185:1000::/home/spark:/bin/bash
+ set -e
```