

Homework: Spark

1. จงนับจำนวนร้านอาหารทั้งหมดที่มี

▼ จงนับจำนวนร้านอาหารทั้งหมดที่มี

```
[ ] wongnai_df.count()
```

283508

2. จงนับจำนวนร้านอาหารในกรุงเทพทั้งหมดที่มี

▼ จงนับจำนวนร้านอาหารในกรุงเทพทั้งหมดที่มี

after looked up for the latitude and longitude that provide, assumed that city_id of Bangkok is 1.0

```
[ ] wongnai_df.filter(wongnai_df['city_id']==1.0).count()
```

90611

3. จงนับจำนวนร้านอาหารญี่ปุ่นในต่างจังหวัด (ที่ไม่ใช่กรุงเทพฯ) ทั้งหมดที่มี

▼ จงนับจำนวนร้านอาหารญี่ปุ่นในต่างจังหวัด (ที่ไม่ใช่กรุงเทพฯ) ทั้งหมดที่มี

```
[ ] from pyspark.sql.functions import udf  
    from pyspark.sql.types import StringType
```

```
[ ] def category_mapping(cat_id):  
    if cat_id in cat_mapping:  
        return cat_mapping[cat_id]  
    else:  
        return 'Unknown'
```

```
    to_category = udf(category_mapping, StringType())
```

```
[ ] res_outside_bkk = wongnai_df.filter(wongnai_df['city_id']!=1.0)  
    res_cat = res_outside_bkk.withColumn('category', to_category(res_outside_bkk.category_id))
```

```
[ ] jap_res_outside_bkk = res_cat.filter(res_cat['category']=='Japanese')  
    jap_res_outside_bkk.count()
```

3053

4. จงแสดงรายชื่อร้านอาหารที่มีจำนวนการ check-in มากกว่า 300 ครั้ง

▼ จงแสดงรายชื่อร้านอาหารที่มีจำนวนการ check-in มากกว่า 300 ครั้ง

```
[ ] wongnai_df.filter(wongnai_df['number_of_checkins'] > 300).select('name').show()
```

```
+-----+
|      name|
+-----+
| เจ็โรว ข้าวต้มเป็ด|
|      มมอรรอย|
|      มนคันมสด|
|      อบอรรอย|
|      ต้อง เต็ม โต๊ะ|
|      The Glass House|
| Annyeong Korean BBQ|
|      กล้วยน้ำว่า|
|      สวนผัก โอ้กะจู้ อ...|
|      ไม้ออก|
+-----+
```

5. จงหาค่าเฉลี่ย (mean) ของราคาเฉลี่ย (avg_price) ของร้านอาหารทั้งหมด

▼ จงหาค่าเฉลี่ย (mean) ของราคาเฉลี่ย (avg_price) ของร้านอาหารทั้งหมด

```
[ ] from pyspark.sql.functions import mean
wongnai_df.select(mean('avg_price')).show()
```

```
+-----+
| avg(avg_price)|
+-----+
|326.0524635946539|
+-----+
```