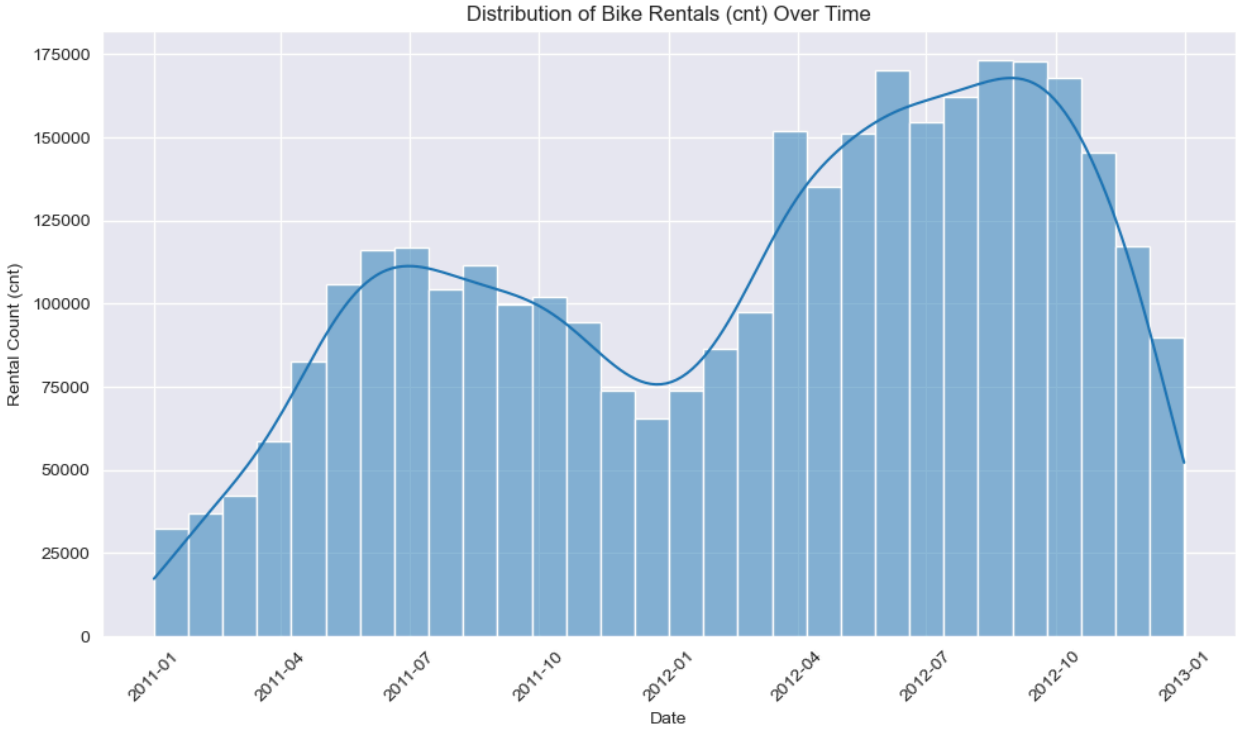
基于时间序列与机器学习的共享单车需求预测

```
pandas~=2.2.2
numpy~=1.26.4
matplotlib~=3.9.2
seaborn~=0.13.2
scikit-learn~=1.5.1
```

首先对数据进行分析：



如图，自行车租赁量与时间的分布

自行车租赁量和体感温度/风速的关系。这里使用了seaborn库，比`matplotlib`更美观，同时能够体现箱线图的部分特征。

随后对数据训练和预测做准备：

```
categorical_cols = ['season', 'mnth', 'hr', 'weekday',
'weathersit'] #对数据进行热编码
```

```
df_encoded['lag1'] = df_encoded['cnt'].shift(1)
df_encoded['lag2'] = df_encoded['cnt'].shift(2) # 构造滞后特征
```

随后进行数据集的划分 运用了 `sklearn`中的model_selection 库

```
X_train, X_test, y_train, y_test = train_test_split(X, y,
test_size=0.2, random_state=42)
# 测试集训练集2:8
```

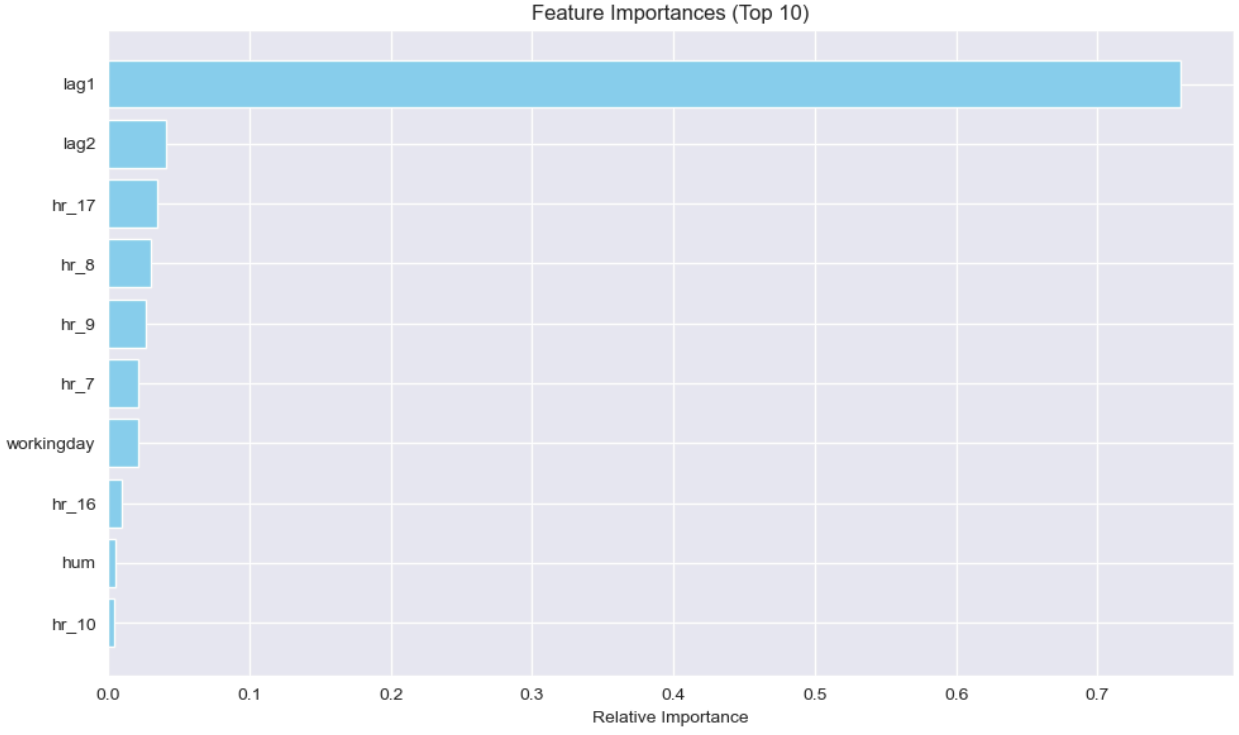
对数据使用了LinearRegression以及RandomForestRegressor进行拟合

```
from sklearn.linear_model import LinearRegression
from sklearn.ensemble import RandomForestRegressor
```

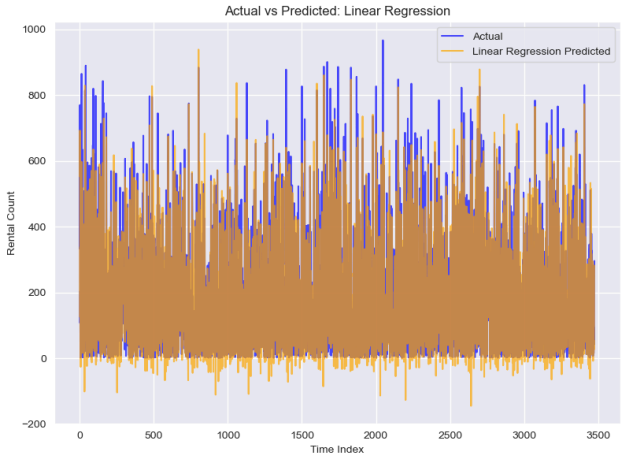
对模型进行了评估，使用到`sklearn.metrics`
结果如下：

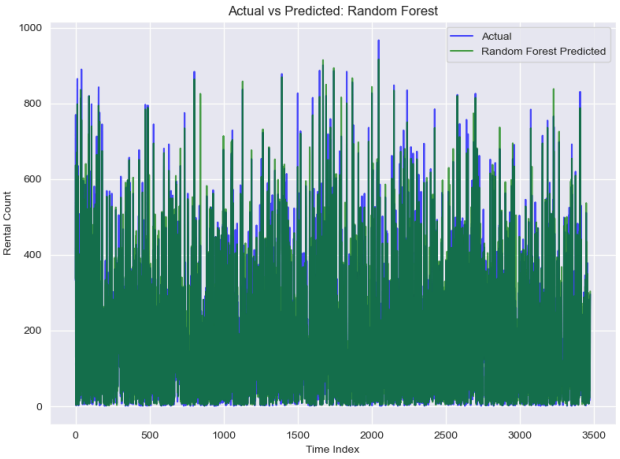
```
Linear Regression:
MSE: 3565.107908470584
R²: 0.8848323418898291

Random Forest Regressor:
MSE: 1192.1726931242808
R² Score: 0.9614879154698799
```

可见`Random Forest Regressor`明显更优。
绘制了对单车租赁量影响最大的前十个特征图。

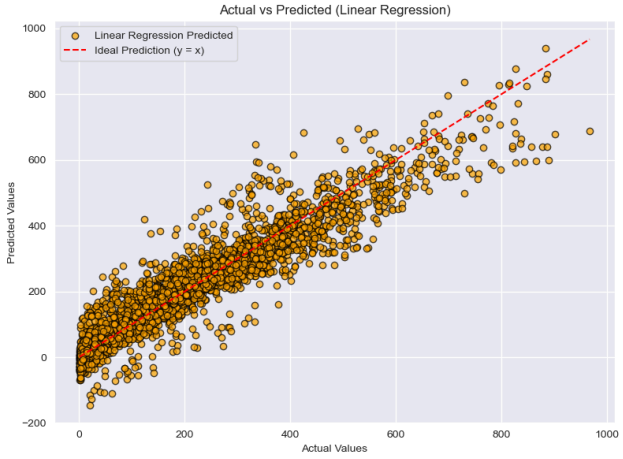


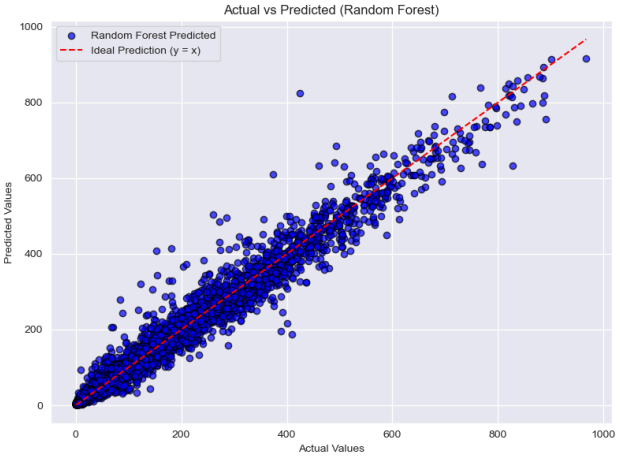
可见，前一个小时的租赁量影响最大，其次是高峰时间段。
最后对模型的优劣进行可视化比较：



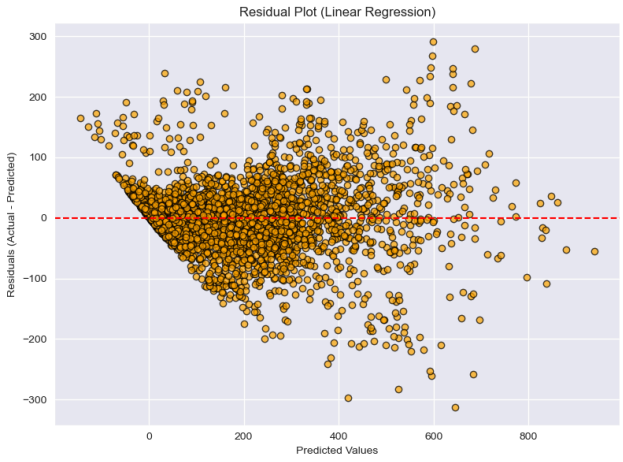


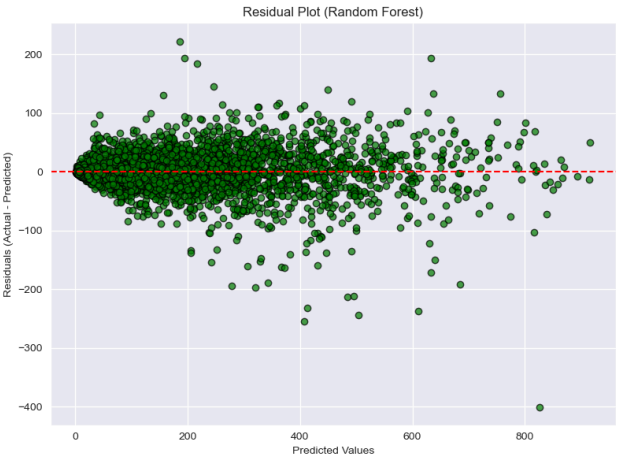
柱状分布不够直观，又绘制了`Actual vs. Predicted Plot`图





以及残差图





可见`Random Forest`模型更优