

Deliverable 3:

Is there a difference in casual bike rentals on works day vs non-workdays

Rebekah Sander





Research Question

Is there a difference in the amount of daily count of casual bike rentals on working days versus days that are not working days?

Research Variable

WHO	WHAT measurement is made on each		TYPE OF MEASURE
	Name of Variable	Question Asked	
A day in 2011 and 2012	Work Day	Is it a work day?	Categorical Variable Unit: working day/not working day
	Casual Counts	How many casual Bike rentals?	Quantitative Variable Unit: Rental Counts
One quantitative variable being tested against one categorical variable to see difference amongst levels.			
1. Two Mean t Test (Pooled) 3. Wilcoxon Rank Sum			
2. Two Mean t Test (Satterthwaite) 4. Welch's Test on Ranked Data			

SAS Code: Renaming



```
/* Make a new data set and RENAMING variables and cat. vars*/  
data work.bike;  
    set work.bike_full (keep = workingday casual);  
    length workingdayC $50;  
    if workingday = 1 then workingdayC = 'working day';  
    else if workingday = 0 then workingdayC = 'not working day';  
    else workingdayC = "Missing";  
    drop workingday;  
    rename workingdayC='Work Day'n;  
    rename casual='Casual Counts'n;  
run;
```

Summary Statistics

- ▶ Mean: The average casual rental count for a non-working day is 1371.13 rentals.
- ▶ Median: 50% of all non-working days have casual rental counts below 1338 rentals.
- ▶ Standard Deviation: On average, one non-working day will have a casual rental count that is off by 873.06 rentals from the mean of 1371.13 rentals.



Analysis Variable : Casual Counts casual					
Work Day	N Obs	Mean	Median	Std Dev	N Miss
not working day	231	1371.13	1338.00	873.06	0
working day	500	606.57	616.50	391.50	0

```
/* Check for missing values and Summary Statistics*/  
proc means data=work.bike mean median stddev nmiss;  
  var 'Casual Counts';  
  class 'Work Day';  
run;
```

Assessing Normality

- ▶ **Sample Size:** There are 231 observations in the not working day group and 500 observations in the working day group. Both samples have sizes over 30 and by the Central Limit Theorem, we may assume normality.
- ▶ **QQ Plot:** Both samples show some deviation in the tails of the agreement line, not enough to reject normality.c



```
/*QQ Plots and normality test*/  
title "Figure 1: QQ Plots";  
proc univariate data=work.bike normaltest plots;  
class 'Work Day';  
VAR 'Casual Counts';  
run;  
title;
```

Figure 1: QQ Plot of Casual Counts for not working day

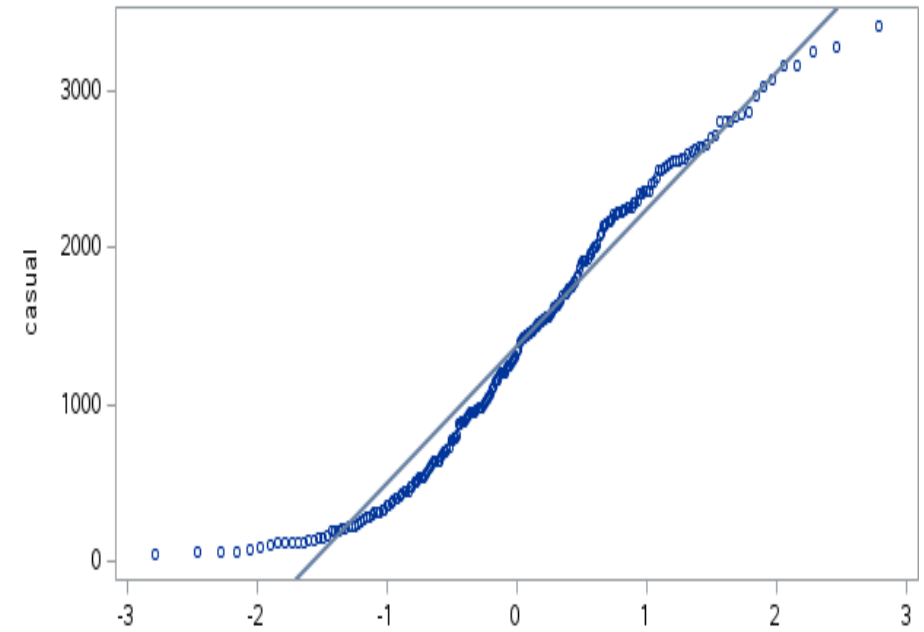
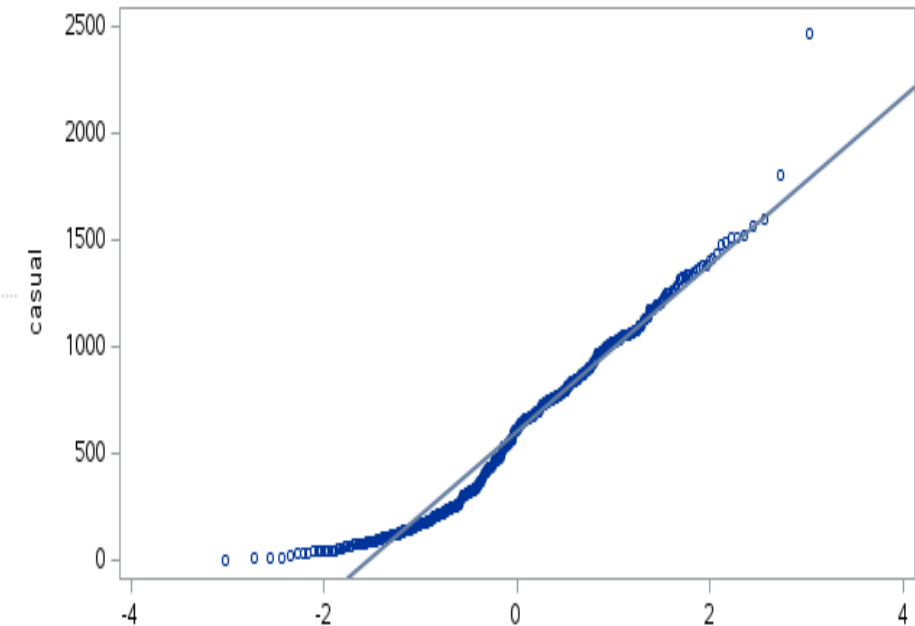


Figure 2: QQ Plot of Casual Counts for working day



Assessing Homogeneity

- ▶ **F-Statistic:** The sample variance in casual rental counts for not working days (873.06^2) is 4.97 times larger than the variance of working days (391.5^2).
- ▶ **P-value:** There is a less than 0.01% chance of getting random samples with these variances when the groups have the same population variances. $\sigma_{not\ WD}^2 = \sigma_{WD}^2$
- ▶ **Conclusion:** At the 0.05 level of significance, we can say that the population variances are different for the casual rental counts between working days and non-working days. Thus, we cannot pool the variances for the t-test.
- ▶ **Appropriate Test:** Satterthwaite



$$H_0: \sigma_{not\ WD}^2 = \sigma_{WD}^2 \text{ (POOL)}$$

$$H_A: \sigma_{not\ WD}^2 \neq \sigma_{WD}^2 \text{ (DON'T POOL)}$$

$$\alpha = 0.05$$

Equality of Variances				
Method	Num DF	Den DF	F Value	Pr > F
Folded F	230	499	4.97	<.0001

Choosing Hypothesis Test: Two mean t-Test (Satterthwaite)

- ▶ Since the data is normal and we do not have equality of variances, we will use the Satterthwaite two mean t-Test.

$$H_0: \mu_{not\ WD} = \mu_{WD}$$

$$H_A: \mu_{not\ WD} \neq \mu_{WD}$$

$$\alpha = 0.05$$

- ▶ The null hypothesis is that the mean rental counts of casual rentals in working days is the same as the mean rental counts of casual rentals in not working days.
- ▶ The alternative hypothesis is that the mean rental counts of casual rentals in working days is different than the mean rental counts of casual rentals in not working days.
- ▶ The level of significance, $\alpha = 0.05$, tells us that 5% of the time we will conclude $\mu_{not\ WD} \neq \mu_{WD}$ when $\mu_{not\ WD} = \mu_{WD}$ is actually true.



Performing Hypothesis Test

- ▶ **t-value:** The sample difference of 764.6 rental counts higher for non-working days as compared to working days is 12.73 standard errors above the hypothesized difference in the population means, which is zero.
- ▶ **p-value:** There is a less than 0.01% chance of observing a sample difference in the means to be greater than or equal to 764.6 rental counts in magnitude when there is no difference in the population mean casual rental counts for the two groups.
- ▶ **Conclusion:** $p\text{-value} = 0.001 < \alpha = 0.05$, At the $\alpha = 0.05$ significance level we have evidence to suggest that the casual rental counts between working days and not working days is different. This is statistically significant.

Work Day	Method	N	Mean	Std Dev	Std Err	Minimum	Maximum
not working day		231	1371.1	873.1	57.4434	54.0000	3410.0
working day		500	606.6	391.5	17.5082	2.0000	2469.0
Diff (1-2)	Pooled		764.6	587.7	46.7551		
Diff (1-2)	Satterthwaite		764.6		60.0524		

Method	Variances	DF	t Value	Pr > t
Pooled	Equal	729	16.35	<.0001
Satterthwaite	Unequal	273.63	12.73	<.0001

```
/*Running the two mean t Test*/  
Proc ttest data=work.bike Order=data plots sides=2 H0=0 alpha=.05;  
  Class 'Work Day';  
  Var 'Casual Counts';  
run;
```



Post hoc test: Confidence Interval

Work Day	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
not working day		1371.1	1258.0	1484.3	873.1	800.0	960.9
working day		606.6	572.2	641.0	391.5	368.6	417.4
Diff (1-2)	Pooled	764.6	672.8	856.4	587.7	559.0	619.5
Diff (1-2)	Satterthwaite	764.6	646.3	882.8			



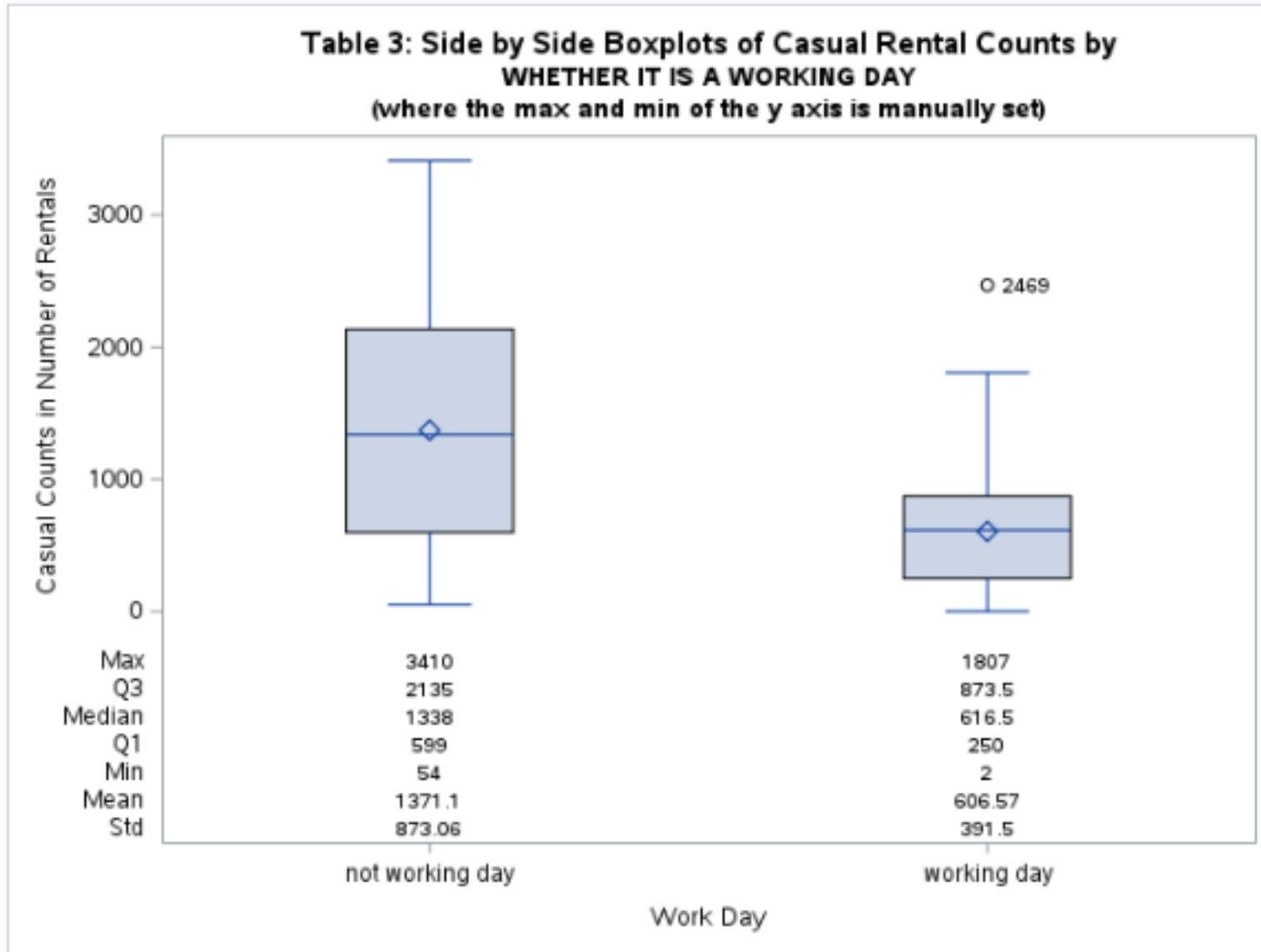
- ▶ We are 95% confident that the true average casual rental count for non-working days is between 646.3 rental counts less to 882.8 rental counts more than working days.
- ▶ Since zero is not included in this interval, there is a statistically significant difference in casual rental counts between working days and not working days.
- ▶ The confidence interval agrees with the Satterthwaite two mean t-test that we can reject H_0 as we have evidence to support H_A

SAS Code: Box Plots

```
5  
6 TITLE1 "Table 3: Side by Side Boxplots of Casual Rental Counts by";  
7 title2 "WHETHER IT IS A WORKING DAY";  
8 title3 "(where the max and min of the y axis is manually set)";  
9 PROC SGPLOT DATA = work.bike;  
0 VBOX 'Casual Counts'n / CATEGORY='Work Day'n boxwidth=0.3 datalabel='Casual Counts'n displaystats=(std MEAN min Q1 MEDIAN Q3 MAX);  
1 YAXIS LABEL = 'Casual Counts in Number of Rentals' MIN=2 MAX=3410;  
2 RUN;  
3 title;
```



Supporting Graphic: Boxplot

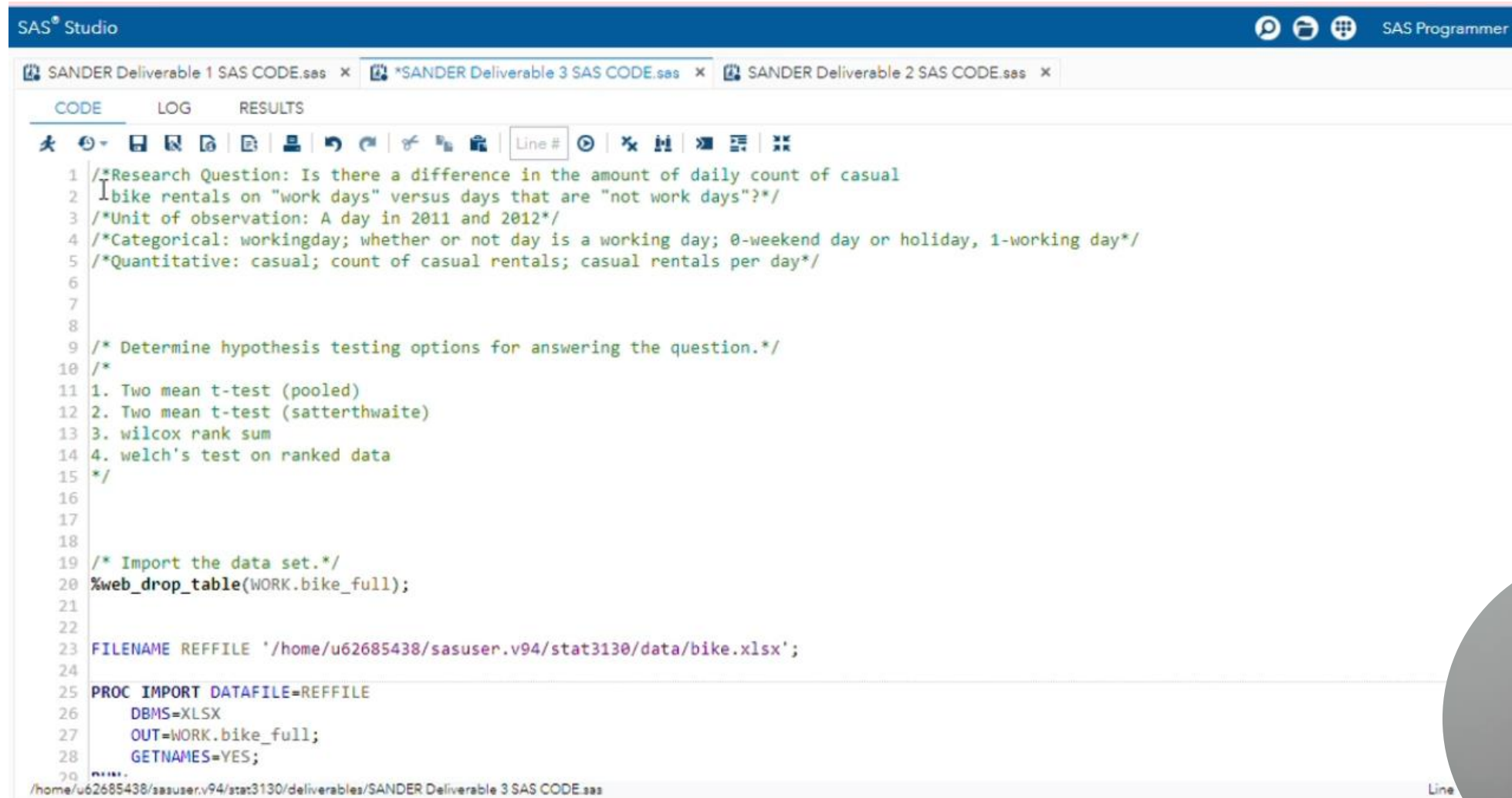


Taking Action

- ▶ This analysis may be helpful to bike rental companies if they are trying to see if they need to make more bikes accessible in more dense areas of business or in more casual places.



SAS Code Screen Recording



The screenshot displays the SAS Studio interface. The top bar shows 'SAS® Studio' and 'SAS Programmer'. Below the top bar, there are three tabs: 'SANDER Deliverable 1 SAS CODE.sas', '*SANDER Deliverable 3 SAS CODE.sas' (which is the active tab), and 'SANDER Deliverable 2 SAS CODE.sas'. The main editor area has three tabs: 'CODE', 'LOG', and 'RESULTS'. The 'CODE' tab is active, showing a SAS script. The script includes comments about the research question, unit of observation, and categorical/quantitative variables. It also lists hypothesis testing options and includes code to import an Excel file named 'bike.xlsx' into a SAS dataset named 'WORK.bike_full'.

```
1 /*Research Question: Is there a difference in the amount of daily count of casual
2  bike rentals on "work days" versus days that are "not work days"?*/
3 /*Unit of observation: A day in 2011 and 2012*/
4 /*Categorical: workingday; whether or not day is a working day; 0-weekend day or holiday, 1-working day*/
5 /*Quantitative: casual; count of casual rentals; casual rentals per day*/
6
7
8
9 /* Determine hypothesis testing options for answering the question.*/
10 /*
11 1. Two mean t-test (pooled)
12 2. Two mean t-test (satterthwaite)
13 3. wilcox rank sum
14 4. welch's test on ranked data
15 */
16
17
18
19 /* Import the data set.*/
20 %web_drop_table(WORK.bike_full);
21
22
23 FILENAME REFFILE '/home/u62685438/sasuser.v94/stat3130/data/bike.xlsx';
24
25 PROC IMPORT DATAFILE=REFFILE
26     DBMS=XLSX
27     OUT=WORK.bike_full;
28     GETNAMES=YES;
29 *****
30
```

The status bar at the bottom shows the file path: '/home/u62685438/sasuser.v94/stat3130/deliverables/SANDER Deliverable 3 SAS CODE.sas'.

