R Report for Cab prediction

Pre-requisites:

1. To run the R project the system shall have:
2. 64 bit R
3. Rstudio
4. Download the file and save it in your local in Documents folder. The file format should be in below format:

C:/Users/USER\_NAME/Documents/train.csv (in my case : C:/Users/Ananya Verma/Documents/train.csv)

AND

C:/Users/USER\_NAME/Documents/test.csv (in my case : C:/Users/Ananya Verma/Documents/test.csv)

Steps:

1. Run the application by following the below steps using Command Prompt:
2. Open command prompt by typing “cmd” in start.
3. Type “cd path-where –R-is-installed” Note that cd and path has a space between them.
4. Next, type the file name in that path and press enter.

We will start building our model by storing our data in test and train objects.

The dimensions of train data is 16067 rows and 7 columns, and test data dimensions are 9914 rows 6 columns.

Finding the first 5 values of train and test data and checking the summary of data, we have few current observations:

* Fare\_amount is not a binary/categorical data.
* Pickup\_datetime is in UTC format having time and date of a particular ride. We will need to treat pickup\_datetime variable in such a way that we can use algorithms with ease
* Fare\_amount is negative, since amount cannot be negative we will remove negative values
* Data has missing values

We investigate more on fare\_amount, we find that there is a value (Other) and holds a value of “430-”. This value doesn’t make any sense hence we delete this value.

train[1124,]

train = train[-c(1124), ]

Using mutate function, we divide the pickup\_datetime into month, hour, Day of Week and year for ease in implementing the regression algorithms in both train and test data. We shall also delete the old variable pickup\_datetime.

We now check the missing values in train and test sets and impute them with median, train set has 60 missing values whereas test set has zero.

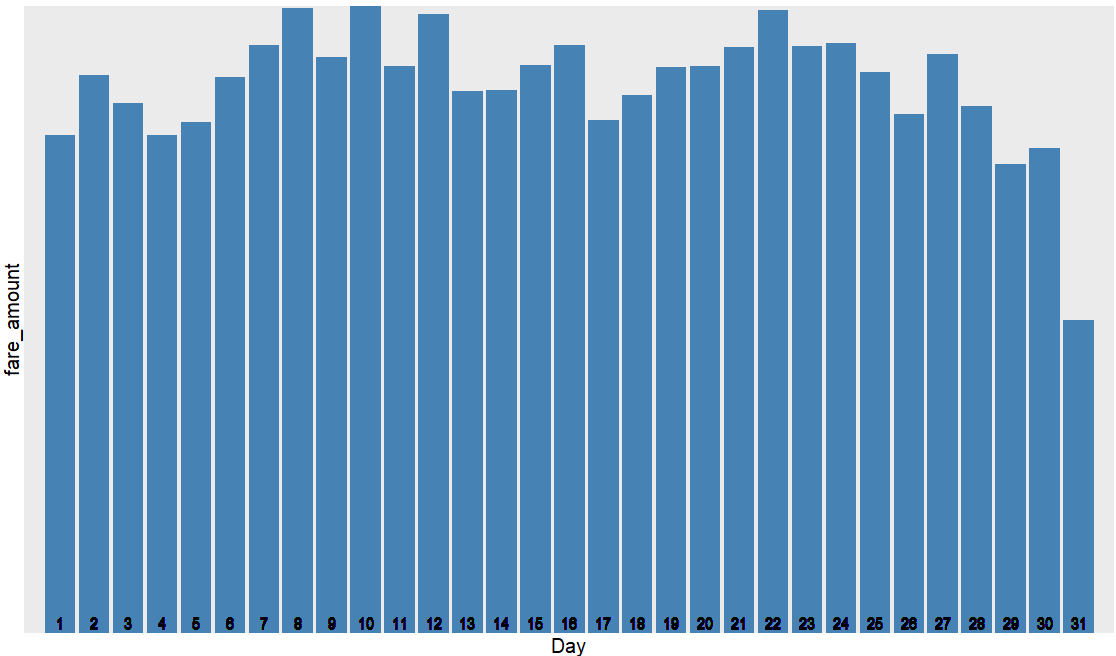
The NA’s in train and test set is zero.

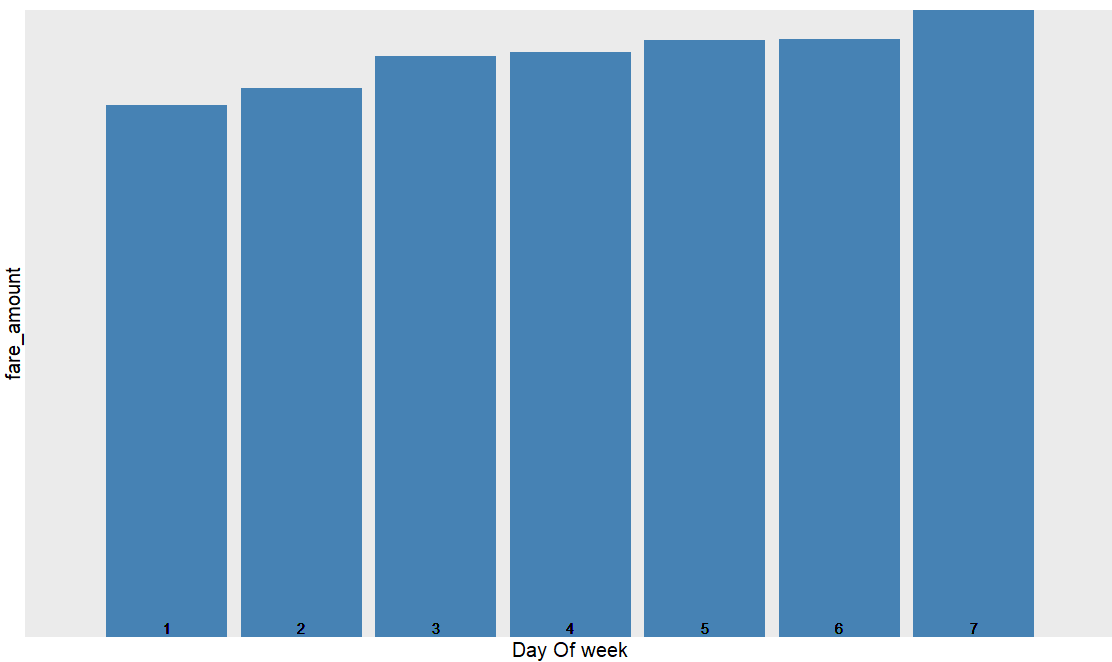
Dealing with Longitude and Latitude data, Lines of latitude range from 0-90 degrees both north and south, while lines of longitude range from 0-180 degrees both east and west. We check that all the values satisfy this condition.

We shall now visually check what data tries to say,

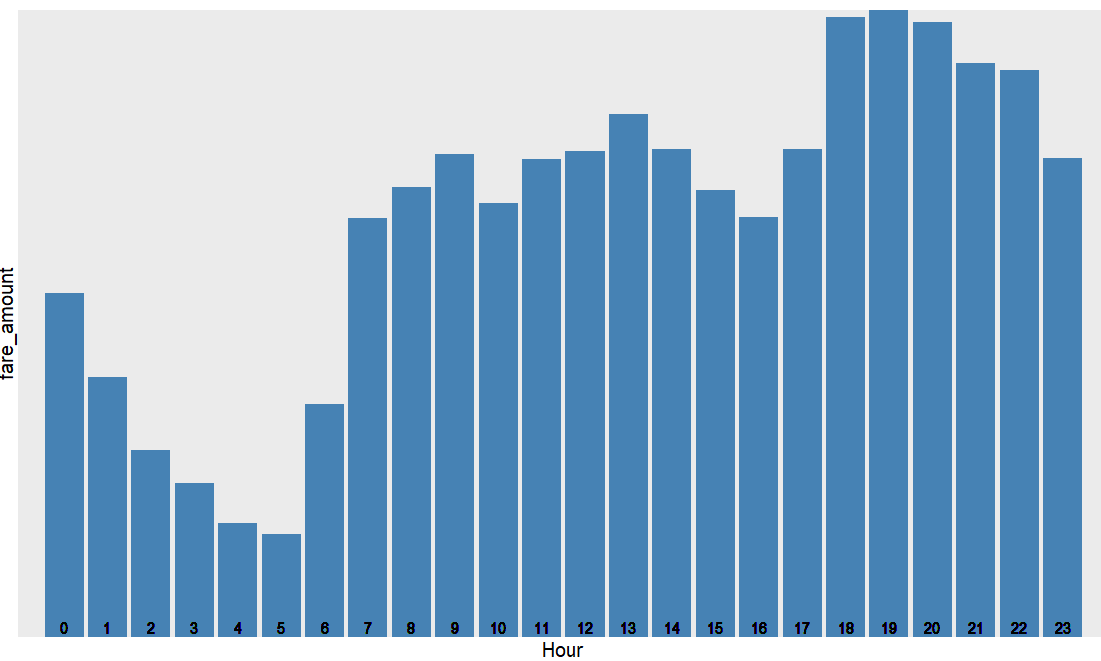
The fare amount was the highest in the 8th,10th and 22 day.

The fare\_amount was the highest on Sunday (considering 1 as Monday)

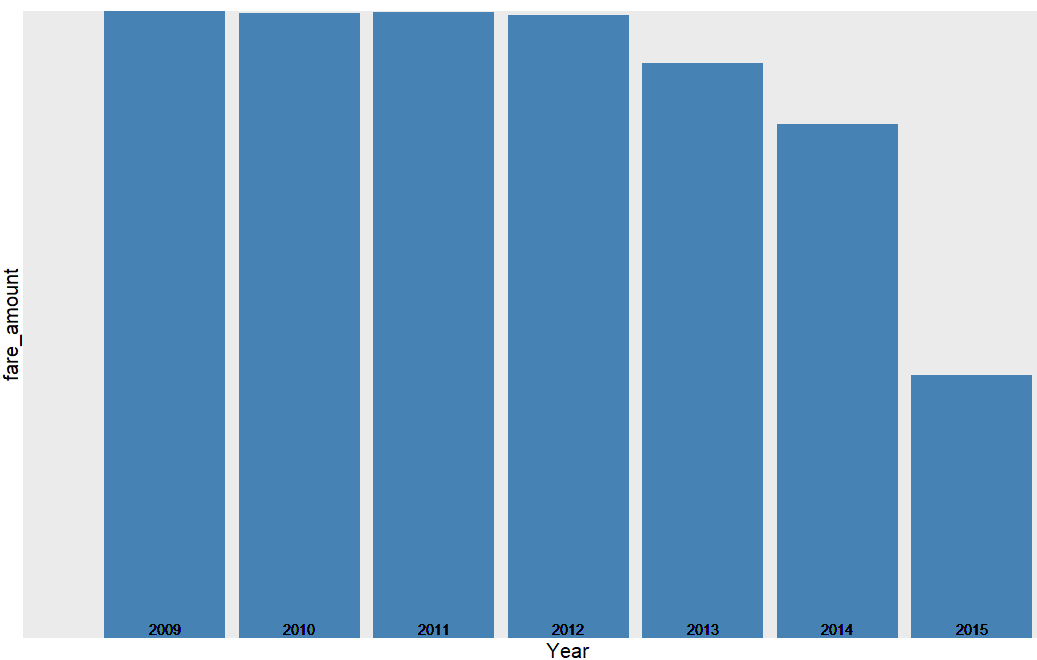




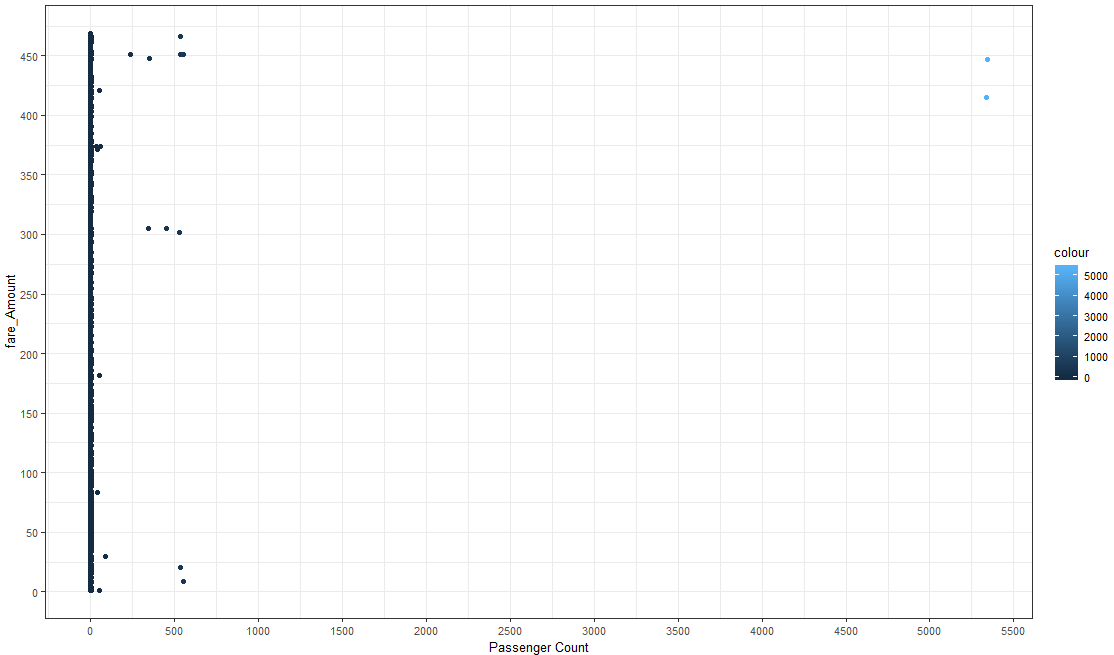
Fare amount was also the highest at 19:00 Hours (or 7:00 PM) may be due to surge in prices due to higher demands, or lesser cabs available at that time.



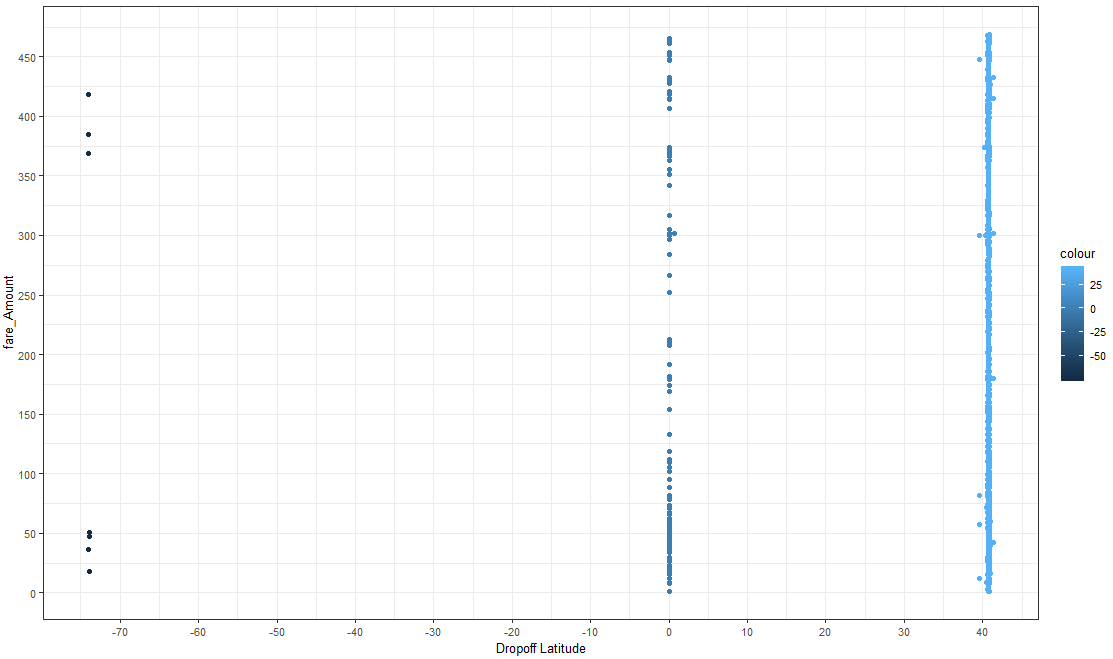
The fare amount was collected the highest in 2009, 2010 2011 and 2012 and later went down in 2015, we are assuming that the company demand was decreased or a competitor had a hold in the industry due to lesser customers.



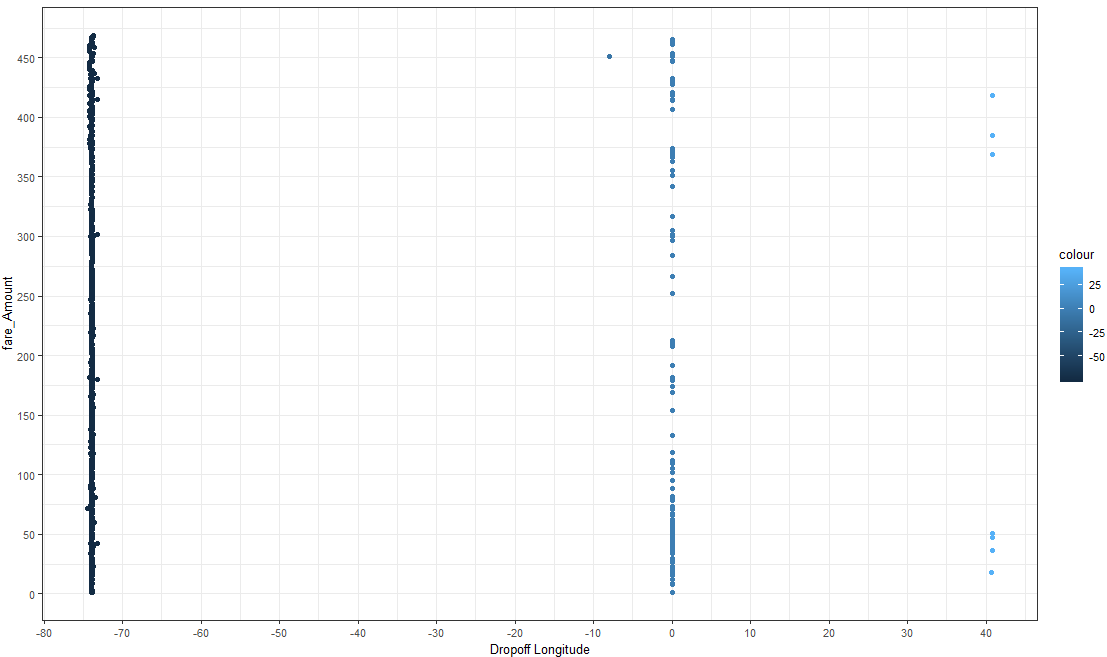
We also find that the passenger count has some outliers which is very evident from the plot below.



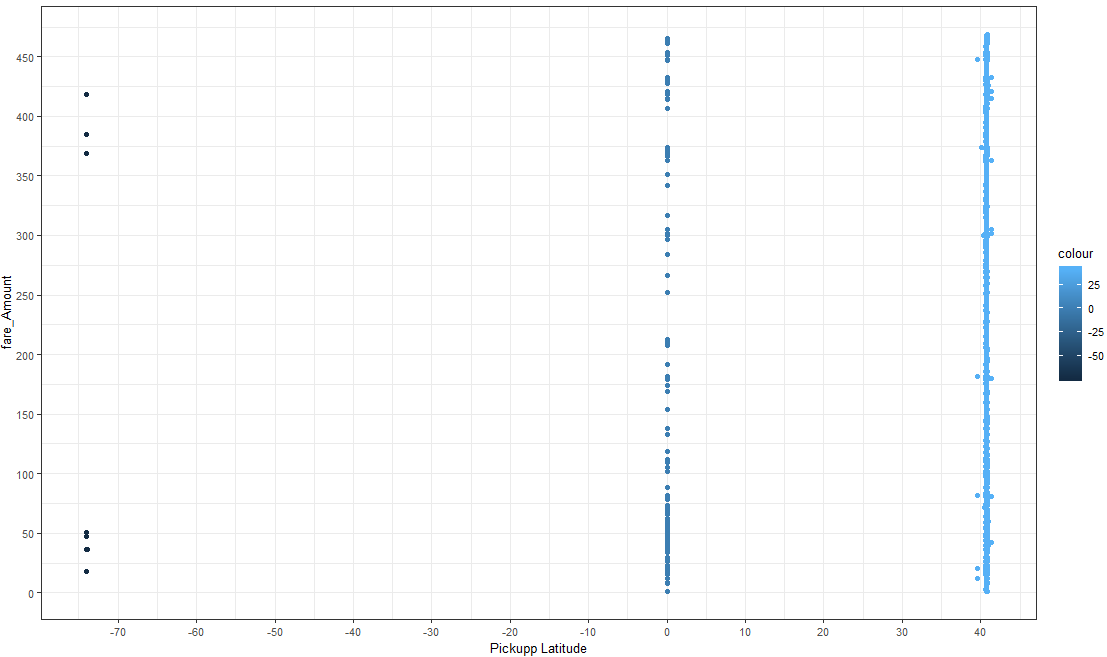
The below plot shows relationship between dropoff latitude and fare amount.



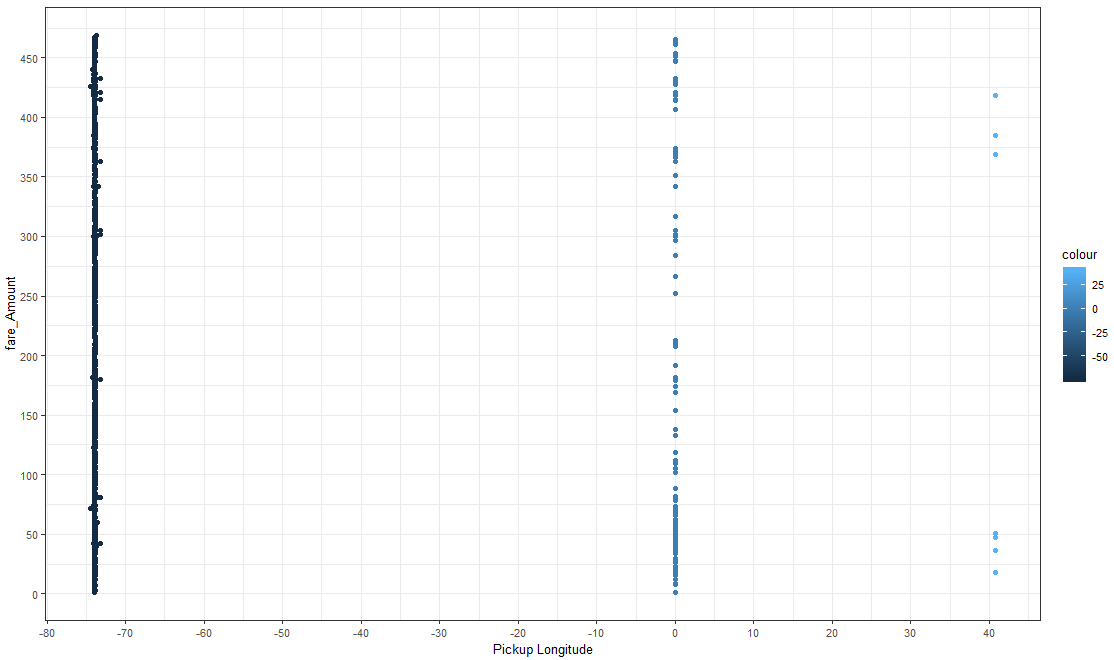
The below plot shows relationship between dropoff longitude and fare amount.



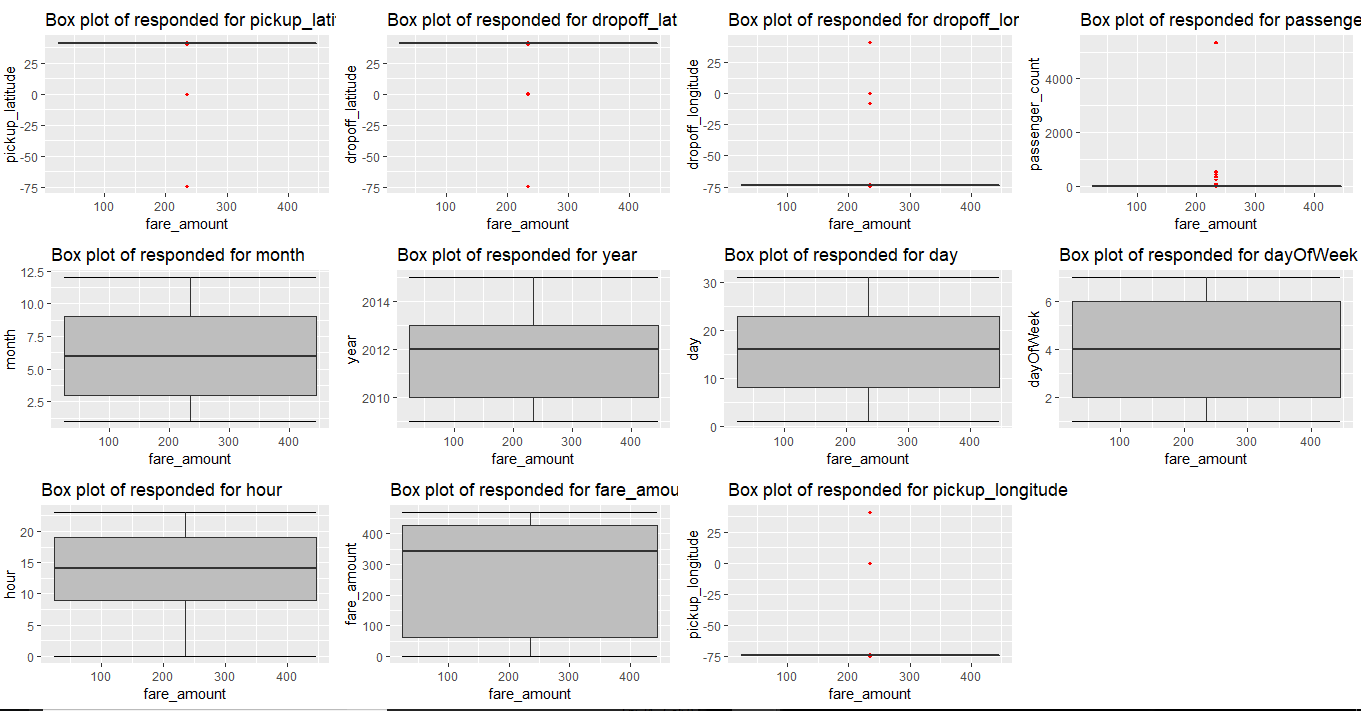
The below plot shows relationship between Pickup latitude and fare amount.



The below plot shows relationship between Pickup longitude and fare amount.

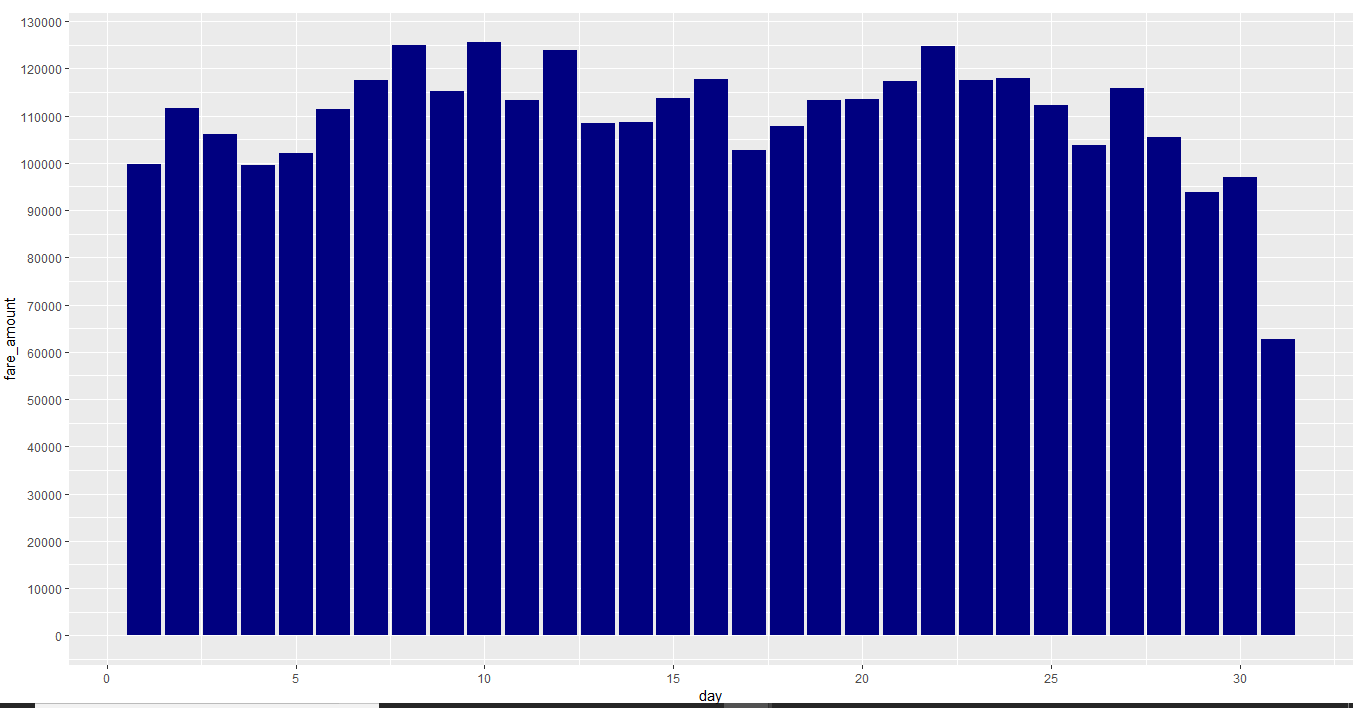


Checking for Outliers:

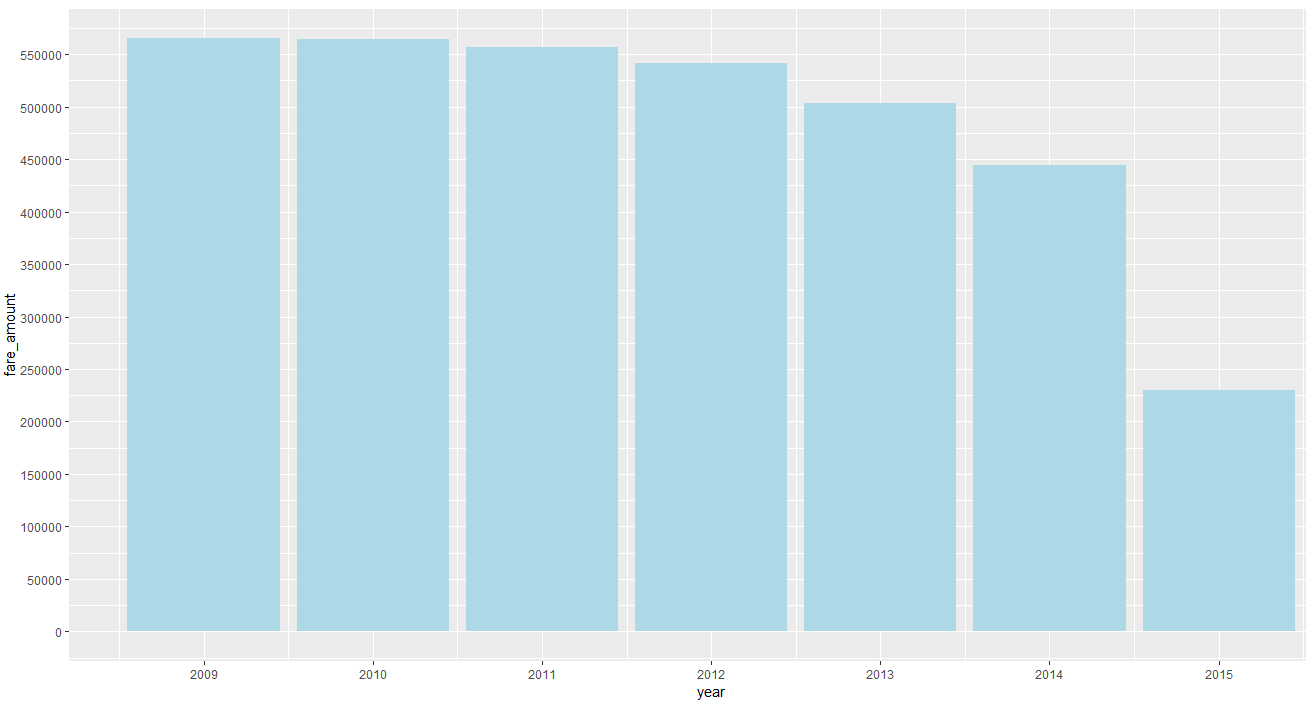


We observe that pickup latitude dropoff latitude dropoff longitude , pickup longitude and passenger count have outliers hence we will remove the outliers.

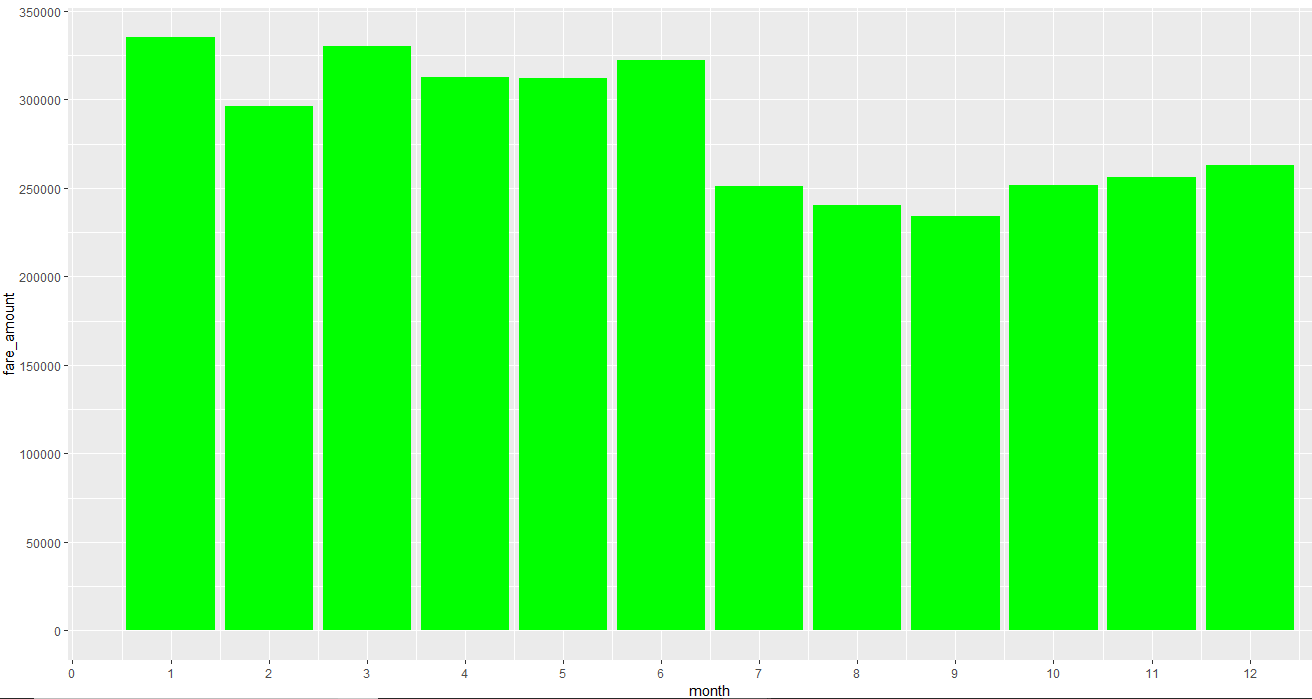
Few observations via plotting after removing outlier.



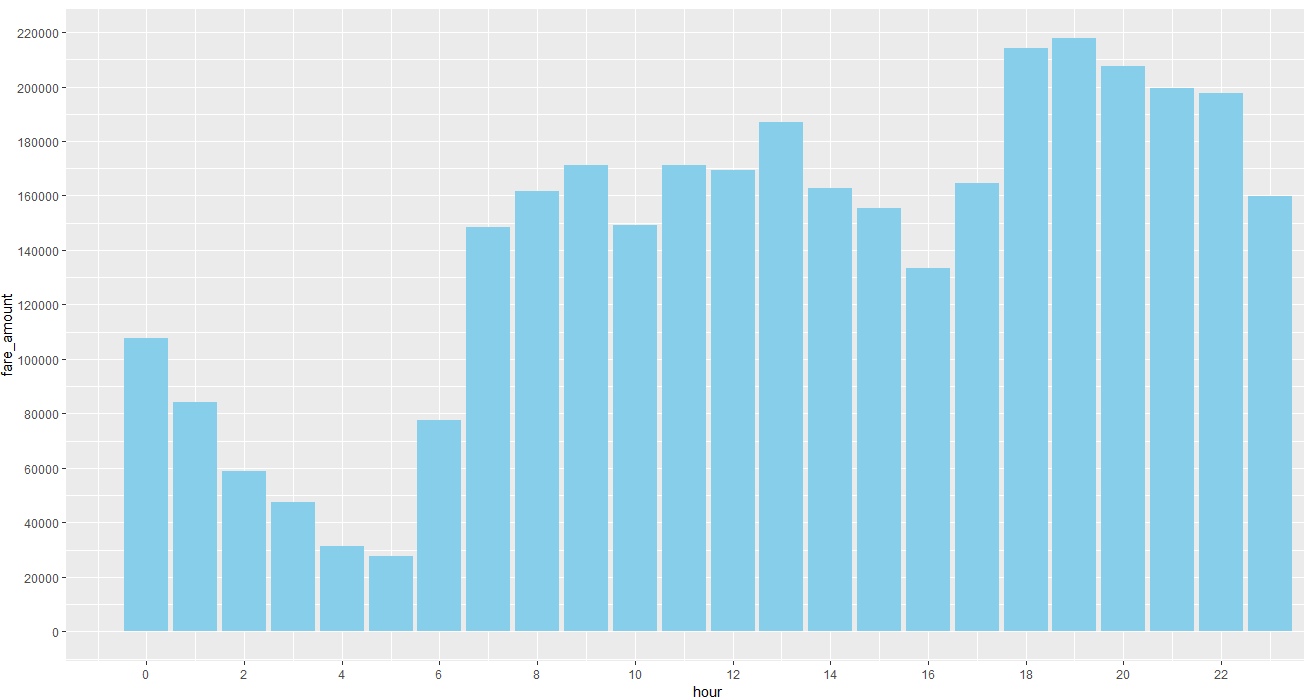
Fare amount was highest in year 2009.

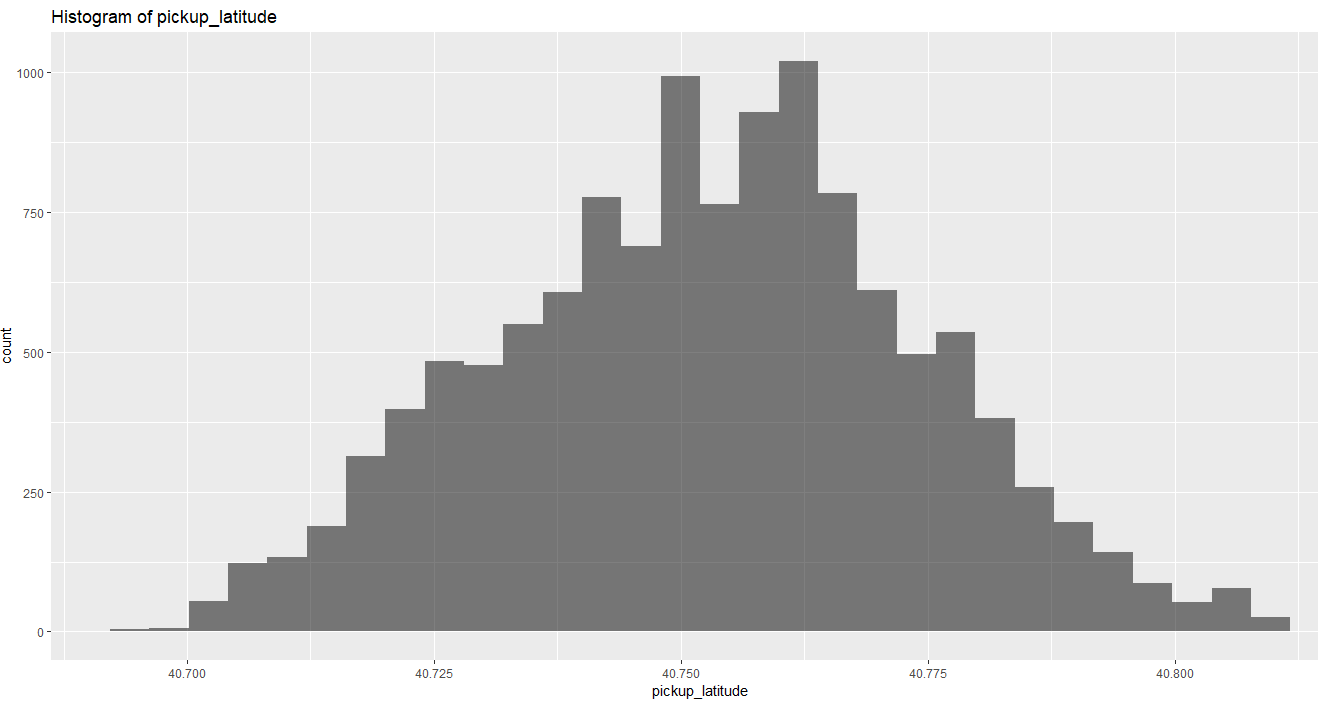


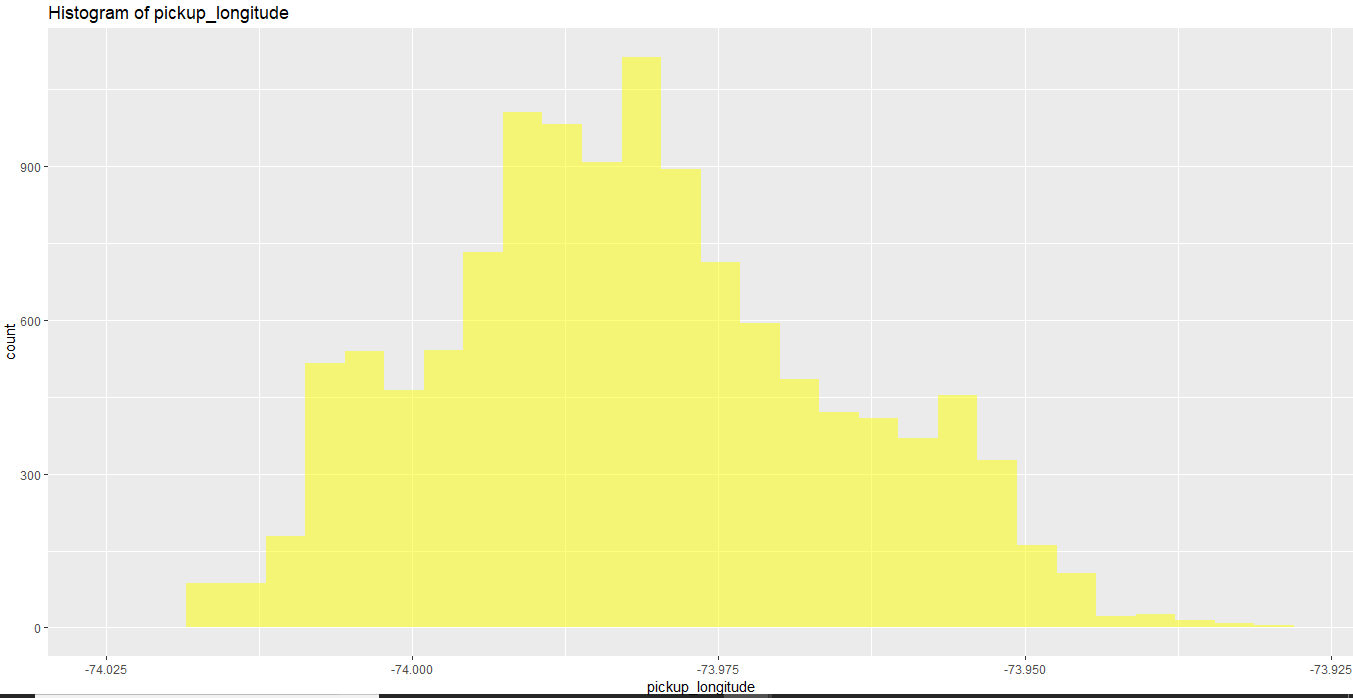
Fare amount was the highest in January.



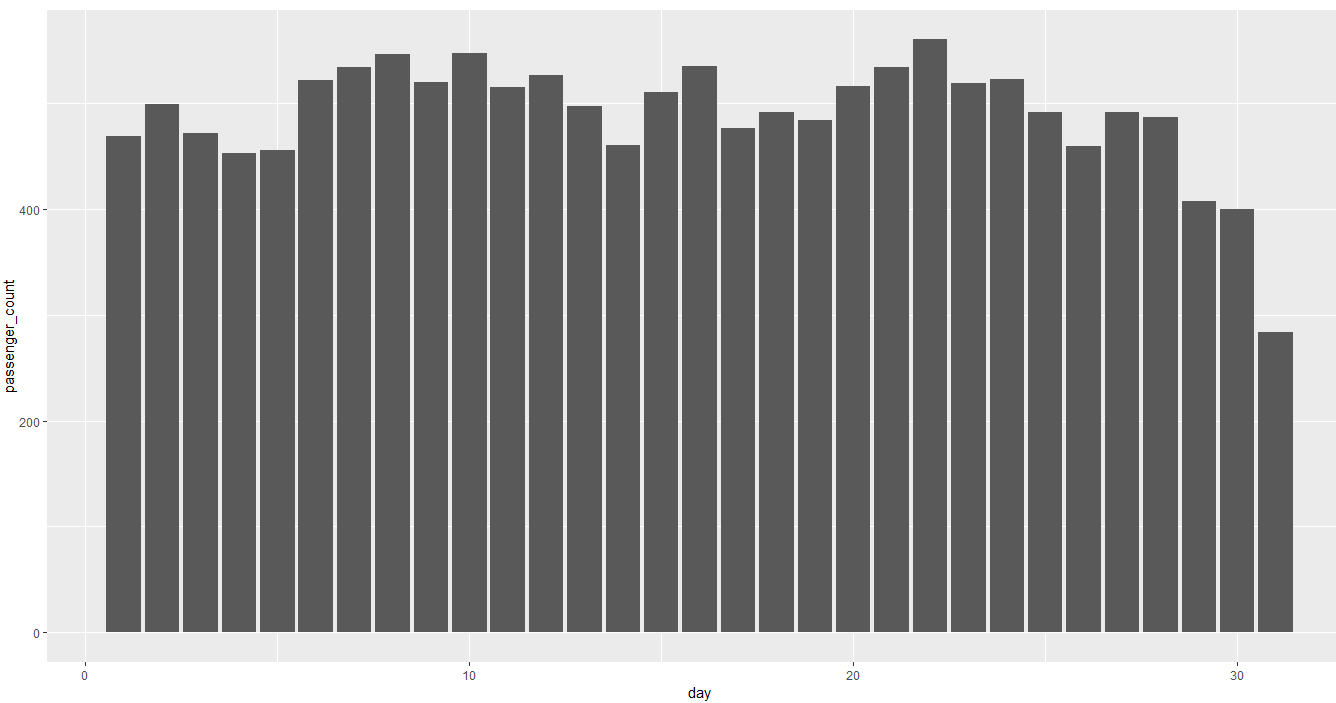
Fare amount came out to be maximum at 19:00 hours, or 7:00 PM.

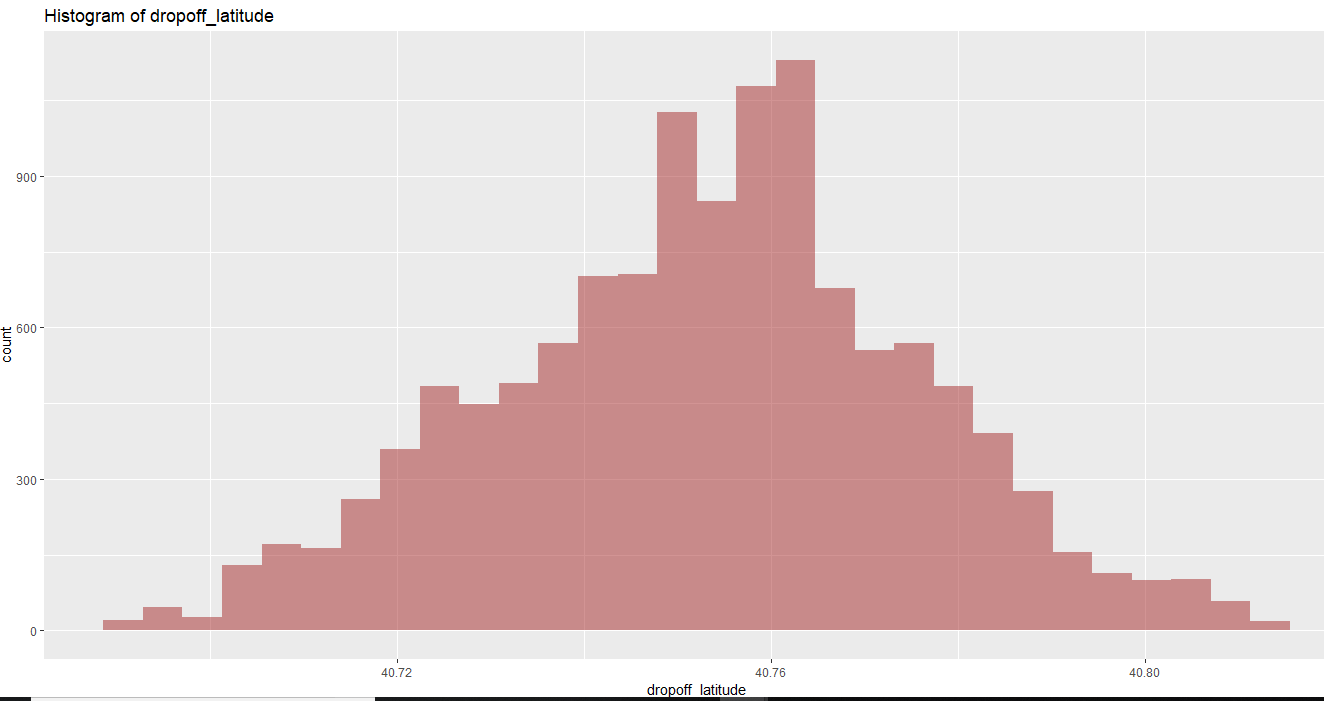




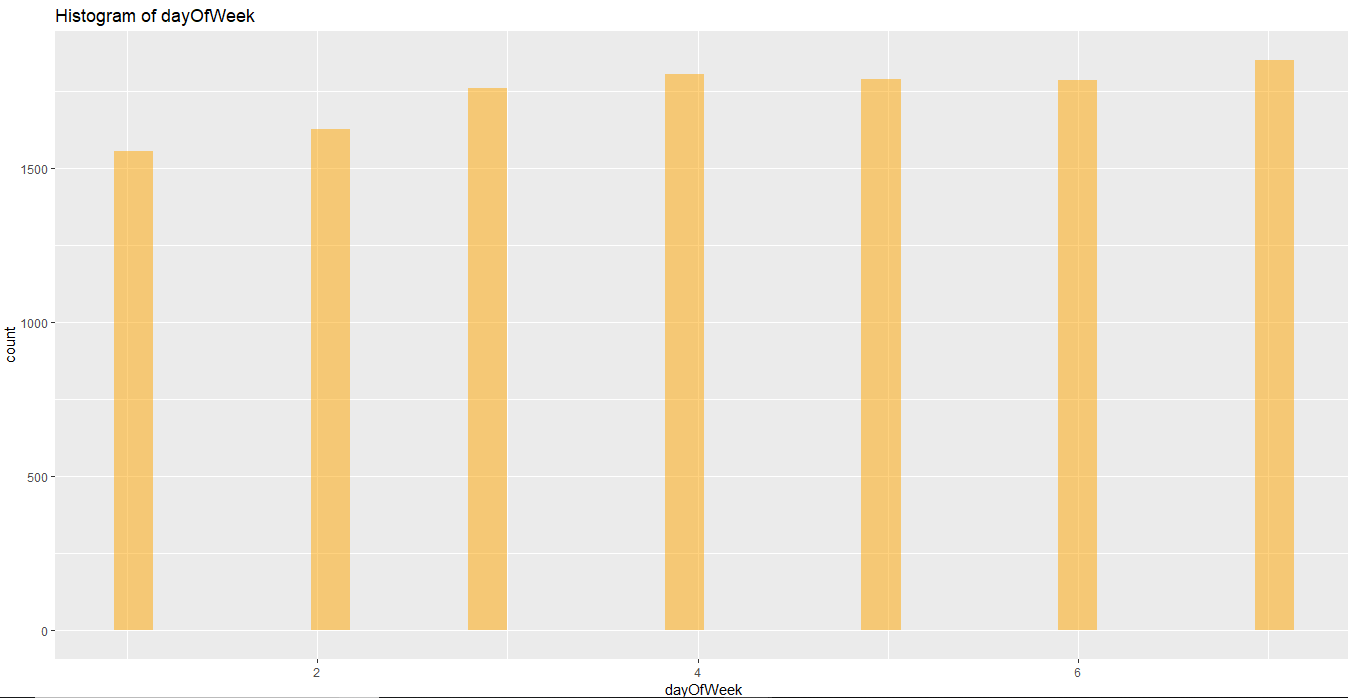


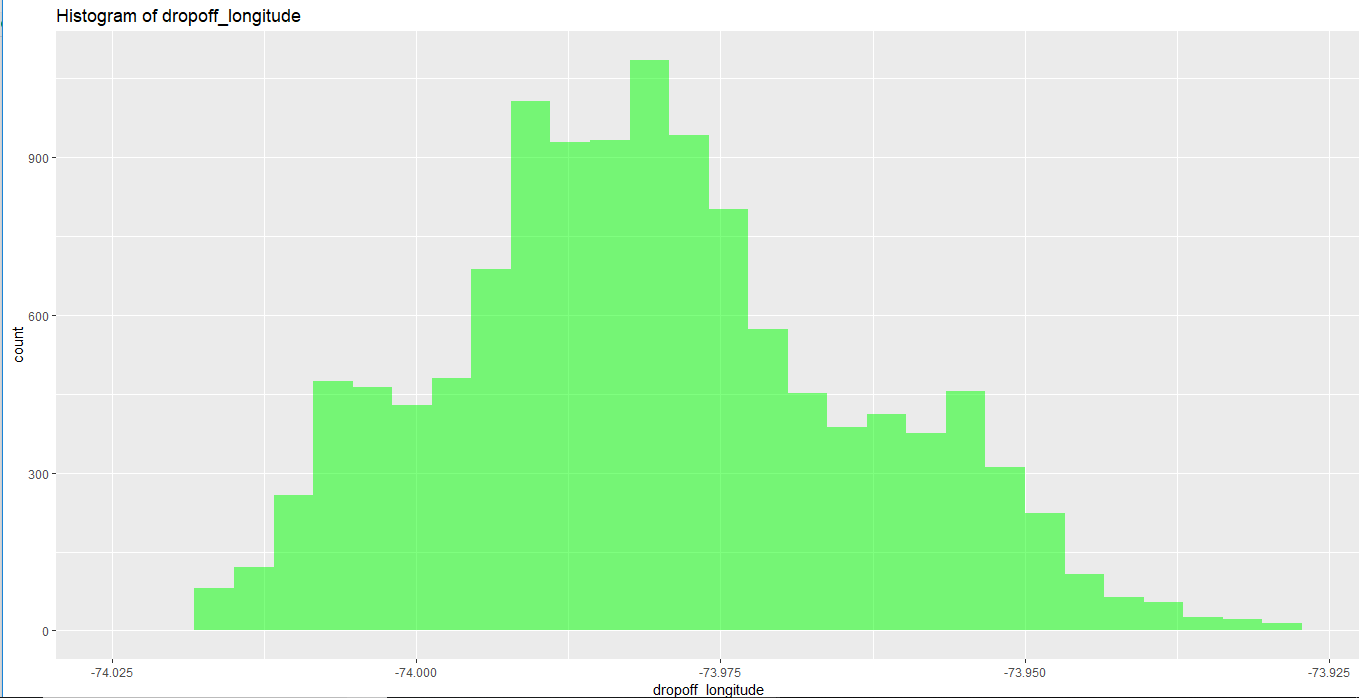
Maximum passengers travelled on 22nd day.





And the cab ride were taken the maximum on Sunday(Assuming 1 as Monday).





We have used Decision tree and linear regression models. We have dropped Logistic regression as we do not need to calculate binary data or probability of an event to occur.

Decision tree observations (Using MAPE error metric). RMSE is not used since RMSE mainly deals with time-series data.

#Error rate : 3.900695

#Accuracy : 96.1

Whereas, linear regression model gave below observations:

#Error rate : 3.926267

#Accuracy : 96.08

We see that the decision tree model gives better accuracy, hence we shall select this model.

1 2 3 4 5 6 7 8

329.66652 329.66652 329.66652 329.66652 128.00000 313.73344 313.73344 75.55022

9 10 11 12 13 14 15 16

260.14155 315.57463 342.38228 121.31354 329.66652 315.57463 329.66652 121.31354

17 18 19 20 21 22 23 24

329.66652 320.47588 329.66652 329.66652 271.23604 320.47588 329.66652 315.57463

25 26 27 28 29 30 31 32

89.21389 315.57463 89.21389 121.31354 320.47588 89.21389 320.47588 329.66652

33 34 35 36 37 38 39 40

320.47588 315.57463 313.73344 329.66652 329.66652 329.66652 329.66652 320.47588

41 42 43 44 45 46 47 48

315.57463 315.57463 75.55022 315.57463 329.66652 313.73344 329.66652 313.73344

49 50 51 52 53 54 55 56

329.66652 329.66652 271.23604 315.57463 271.23604 315.57463 342.38228 128.00000

57 58 59 60 61 62 63 64

320.47588 315.57463 329.66652 128.00000 271.23604 329.66652 329.66652 329.66652

65 66 67 68 69 70 71 72

329.66652 315.57463 315.57463 315.57463 329.66652 329.66652 342.38228 315.57463

73 74 75 76 77 78 79 80

329.66652 329.66652 329.66652 313.73344 271.23604 315.57463 315.57463 313.73344

81 82 83 84 85 86 87 88

260.14155 320.47588 315.57463 329.66652 89.21389 315.57463 329.66652 329.66652

89 90 91 92 93 94 95 96

260.14155 329.66652 89.21389 329.66652 329.66652 315.57463 128.00000 313.73344

97 98 99 100 101 102 103 104

260.14155 113.94087 329.66652 131.85475 128.00000 320.47588 75.55022 329.66652

105 106 107 108 109 110 111 112

329.66652 313.73344 315.57463 342.38228 342.38228 121.31354 271.23604 89.21389

113 114 115 116 117 118 119 120

329.66652 271.23604 315.57463 315.57463 89.21389 315.57463 329.66652 315.57463

121 122 123 124 125 126 127 128

313.73344 329.66652 320.47588 329.66652 313.73344 342.38228 315.57463 329.66652

129 130 131 132 133 134 135 136

342.38228 89.21389 320.47588 320.47588 313.73344 329.66652 342.38228 315.57463

137 138 139 140 141 142 143 144

113.94087 315.57463 121.31354 342.38228 313.73344 260.14155 329.66652 329.66652

145 146 147 148 149 150 151 152

315.57463 329.66652 128.00000 342.38228 320.47588 329.66652 342.38228 329.66652

153 154 155 156 157 158 159 160

329.66652 313.73344 89.21389 329.66652 313.73344 320.47588 121.31354 315.57463

161 162 163 164 165 166 167 168

113.94087 313.73344 313.73344 89.21389 342.38228 320.47588 271.23604 128.00000

169 170 171 172 173 174 175 176

320.47588 329.66652 329.66652 329.66652 113.94087 329.66652 320.47588 329.66652

177 178 179 180 181 182 183 184

342.38228 329.66652 89.21389 320.47588 89.21389 329.66652 75.55022 329.66652

185 186 187 188 189 190 191 192

315.57463 329.66652 342.38228 105.73292 89.21389 329.66652 313.73344 271.23604

193 194 195 196 197 198 199 200

329.66652 315.57463 329.66652 329.66652 329.66652 113.94087 329.66652 329.66652

201 202 203 204 205 206 207 208

313.73344 329.66652 329.66652 313.73344 329.66652 329.66652 271.23604 329.66652

209 210 211 212 213 214 215 216

329.66652 89.21389 329.66652 113.94087 315.57463 329.66652 89.21389 329.66652

217 218 219 220 221 222 223 224

320.47588 260.14155 271.23604 315.57463 89.21389 329.66652 329.66652 128.00000

225 226 227 228 229 230 231 232

313.73344 329.66652 329.66652 329.66652 329.66652 329.66652 271.23604 329.66652

233 234 235 236 237 238 239 240

329.66652 128.00000 113.94087 315.57463 342.38228 315.57463 320.47588 113.94087

241 242 243 244 245 246 247 248

89.21389 329.66652 315.57463 320.47588 315.57463 128.00000 315.57463 313.73344

249 250 251 252 253 254 255 256

75.55022 329.66652 315.57463 313.73344 113.94087 105.73292 260.14155 121.31354

257 258 259 260 261 262 263 264

329.66652 128.00000 271.23604 271.23604 320.47588 271.23604 89.21389 121.31354

265 266 267 268 269 270 271 272

75.55022 271.23604 89.21389 89.21389 75.55022 320.47588 329.66652 315.57463

273 274 275 276 277 278 279 280

320.47588 128.00000 329.66652 320.47588 315.57463 271.23604 329.66652 315.57463

281 282 283 284 285 286 287 288

128.00000 89.21389 329.66652 89.21389 271.23604 329.66652 329.66652 313.73344

289 290 291 292 293 294 295 296

329.66652 329.66652 121.31354 271.23604 329.66652 320.47588 320.47588 313.73344

297 298 299 300 301 302 303 304

313.73344 329.66652 128.00000 329.66652 329.66652 75.55022 75.55022 75.55022

305 306 307 308 309 310 311 312

320.47588 329.66652 315.57463 131.85475 329.66652 320.47588 329.66652 329.66652

313 314 315 316 317 318 319 320

113.94087 313.73344 342.38228 313.73344 315.57463 315.57463 329.66652 113.94087

321 322 323 324 325 326 327 328

329.66652 329.66652 320.47588 315.57463 313.73344 128.00000 320.47588 315.57463

329 330 331 332 333 334 335 336

315.57463 271.23604 121.31354 121.31354 320.47588 329.66652 313.73344 320.47588

337 338 339 340 341 342 343 344

320.47588 315.57463 313.73344 313.73344 329.66652 313.73344 128.00000 329.66652

345 346 347 348 349 350 351 352

329.66652 329.66652 329.66652 315.57463 320.47588 329.66652 329.66652 320.47588

353 354 355 356 357 358 359 360

271.23604 113.94087 329.66652 329.66652 320.47588 313.73344 75.55022 131.85475

361 362 363 364 365 366 367 368

320.47588 329.66652 329.66652 329.66652 121.31354 315.57463 320.47588 313.73344

369 370 371 372 373 374 375 376

121.31354 315.57463 329.66652 342.38228 89.21389 329.66652 315.57463 320.47588

377 378 379 380 381 382 383 384

329.66652 313.73344 320.47588 313.73344 329.66652 271.23604 329.66652 315.57463

385 386 387 388 389 390 391 392

342.38228 320.47588 313.73344 315.57463 89.21389 313.73344 313.73344 320.47588

393 394 395 396 397 398 399 400

320.47588 315.57463 271.23604 329.66652 320.47588 329.66652 128.00000 113.94087

401 402 403 404 405 406 407 408

315.57463 329.66652 315.57463 121.31354 329.66652 329.66652 313.73344 121.31354

409 410 411 412 413 414 415 416

121.31354 75.55022 313.73344 329.66652 121.31354 271.23604 315.57463 313.73344

417 418 419 420 421 422 423 424

329.66652 313.73344 329.66652 128.00000 329.66652 320.47588 315.57463 320.47588

425 426 427 428 429 430 431 432

320.47588 271.23604 315.57463 313.73344 320.47588 329.66652 329.66652 329.66652

433 434 435 436 437 438 439 440

313.73344 128.00000 313.73344 128.00000 315.57463 329.66652 313.73344 315.57463

441 442 443 444 445 446 447 448

315.57463 260.14155 313.73344 329.66652 313.73344 315.57463 315.57463 329.66652

449 450 451 452 453 454 455 456

329.66652 315.57463 271.23604 313.73344 329.66652 315.57463 121.31354 313.73344

457 458 459 460 461 462 463 464

315.57463 271.23604 329.66652 329.66652 329.66652 329.66652 315.57463 329.66652

465 466 467 468 469 470 471 472

329.66652 329.66652 131.85475 313.73344 89.21389 313.73344 315.57463 320.47588

473 474 475 476 477 478 479 480

329.66652 315.57463 320.47588 89.21389 329.66652 320.47588 320.47588 315.57463

481 482 483 484 485 486 487 488

329.66652 315.57463 342.38228 329.66652 329.66652 131.85475 329.66652 329.66652

489 490 491 492 493 494 495 496

315.57463 260.14155 128.00000 320.47588 315.57463 320.47588 271.23604 320.47588

497 498 499 500 501 502 503 504

329.66652 315.57463 320.47588 329.66652 105.73292 329.66652 128.00000 89.21389

505 506 507 508 509 510 511 512

315.57463 329.66652 313.73344 329.66652 329.66652 121.31354 313.73344 329.66652

513 514 515 516 517 518 519 520

128.00000 329.66652 75.55022 329.66652 313.73344 315.57463 329.66652 329.66652

521 522 523 524 525 526 527 528

329.66652 128.00000 105.73292 329.66652 329.66652 315.57463 128.00000 315.57463

529 530 531 532 533 534 535 536

313.73344 315.57463 89.21389 131.85475 329.66652 128.00000 315.57463 342.38228

537 538 539 540 541 542 543 544

329.66652 329.66652 329.66652 329.66652 315.57463 329.66652 271.23604 320.47588

545 546 547 548 549 550 551 552

329.66652 315.57463 89.21389 313.73344 329.66652 315.57463 315.57463 89.21389

553 554 555 556 557 558 559 560

342.38228 320.47588 313.73344 271.23604 260.14155 113.94087 89.21389 128.00000

561 562 563 564 565 566 567 568

329.66652 271.23604 320.47588 329.66652 315.57463 313.73344 128.00000 260.14155

569 570 571 572 573 574 575 576

329.66652 329.66652 329.66652 329.66652 329.66652 329.66652 342.38228 320.47588

577 578 579 580 581 582 583 584

260.14155 89.21389 329.66652 329.66652 313.73344 271.23604 313.73344 329.66652

585 586 587 588 589 590 591 592

313.73344 320.47588 329.66652 313.73344 121.31354 315.57463 128.00000 320.47588

593 594 595 596 597 598 599 600

329.66652 320.47588 113.94087 329.66652 105.73292 320.47588 113.94087 89.21389

601 602 603 604 605 606 607 608

329.66652 329.66652 320.47588 121.31354 329.66652 89.21389 329.66652 89.21389

609 610 611 612 613 614 615 616

113.94087 320.47588 271.23604 329.66652 105.73292 329.66652 329.66652 313.73344

617 618 619 620 621 622 623 624

320.47588 315.57463 89.21389 315.57463 89.21389 329.66652 329.66652 329.66652

625 626 627 628 629 630 631 632

121.31354 313.73344 121.31354 329.66652 260.14155 329.66652 75.55022 315.57463

633 634 635 636 637 638 639 640

313.73344 329.66652 329.66652 128.00000 121.31354 113.94087 329.66652 121.31354

641 642 643 644 645 646 647 648

89.21389 329.66652 89.21389 89.21389 329.66652 315.57463 313.73344 313.73344

649 650 651 652 653 654 655 656

315.57463 315.57463 329.66652 329.66652 315.57463 329.66652 329.66652 128.00000

657 658 659 660 661 662 663 664

131.85475 315.57463 128.00000 315.57463 313.73344 128.00000 329.66652 329.66652

665 666 667 668 669 670 671 672

342.38228 271.23604 329.66652 329.66652 342.38228 320.47588 128.00000 329.66652

673 674 675 676 677 678 679 680

315.57463 329.66652 89.21389 329.66652 75.55022 313.73344 315.57463 329.66652

681 682 683 684 685 686 687 688

313.73344 313.73344 329.66652 89.21389 329.66652 320.47588 315.57463 329.66652

689 690 691 692 693 694 695 696

121.31354 313.73344 313.73344 315.57463 313.73344 320.47588 315.57463 75.55022

697 698 699 700 701 702 703 704

329.66652 329.66652 329.66652 329.66652 329.66652 89.21389 131.85475 313.73344

705 706 707 708 709 710 711 712

329.66652 271.23604 75.55022 131.85475 315.57463 329.66652 315.57463 329.66652

713 714 715 716 717 718 719 720

313.73344 329.66652 113.94087 320.47588 128.00000 320.47588 89.21389 313.73344

721 722 723 724 725 726 727 728

313.73344 315.57463 313.73344 320.47588 342.38228 260.14155 320.47588 113.94087

729 730 731 732 733 734 735 736

313.73344 329.66652 315.57463 320.47588 329.66652 329.66652 113.94087 329.66652

737 738 739 740 741 742 743 744

271.23604 313.73344 315.57463 121.31354 113.94087 313.73344 131.85475 329.66652

745 746 747 748 749 750 751 752

113.94087 329.66652 315.57463 320.47588 313.73344 131.85475 320.47588 315.57463

753 754 755 756 757 758 759 760

89.21389 128.00000 342.38228 320.47588 329.66652 313.73344 313.73344 329.66652

761 762 763 764 765 766 767 768

315.57463 315.57463 315.57463 329.66652 75.55022 329.66652 89.21389 313.73344

769 770 771 772 773 774 775 776

329.66652 75.55022 315.57463 315.57463 128.00000 329.66652 342.38228 313.73344

777 778 779 780 781 782 783 784

313.73344 329.66652 329.66652 329.66652 329.66652 329.66652 320.47588 315.57463

785 786 787 788 789 790 791 792

329.66652 313.73344 329.66652 320.47588 320.47588 329.66652 329.66652 75.55022

793 794 795 796 797 798 799 800

131.85475 329.66652 313.73344 271.23604 329.66652 320.47588 329.66652 329.66652

801 802 803 804 805 806 807 808

329.66652 329.66652 315.57463 128.00000 313.73344 313.73344 313.73344 313.73344

809 810 811 812 813 814 815 816

315.57463 131.85475 128.00000 128.00000 131.85475 313.73344 315.57463 113.94087

817 818 819 820 821 822 823 824

313.73344 342.38228 329.66652 89.21389 329.66652 271.23604 320.47588 128.00000

825 826 827 828 829 830 831 832

313.73344 313.73344 128.00000 75.55022 329.66652 329.66652 260.14155 89.21389

833 834 835 836 837 838 839 840

315.57463 320.47588 320.47588 75.55022 313.73344 315.57463 329.66652 315.57463

841 842 843 844 845 846 847 848

315.57463 315.57463 271.23604 315.57463 329.66652 128.00000 260.14155 89.21389

849 850 851 852 853 854 855 856

128.00000 320.47588 329.66652 329.66652 329.66652 329.66652 329.66652 329.66652

857 858 859 860 861 862 863 864

313.73344 315.57463 329.66652 315.57463 329.66652 113.94087 315.57463 271.23604

865 866 867 868 869 870 871 872

329.66652 329.66652 260.14155 313.73344 329.66652 329.66652 313.73344 75.55022

873 874 875 876 877 878 879 880

329.66652 329.66652 320.47588 329.66652 329.66652 128.00000 329.66652 105.73292

881 882 883 884 885 886 887 888

329.66652 320.47588 313.73344 315.57463 313.73344 315.57463 329.66652 315.57463

889 890 891 892 893 894 895 896

313.73344 121.31354 329.66652 128.00000 342.38228 315.57463 320.47588 329.66652

897 898 899 900 901 902 903 904

260.14155 329.66652 271.23604 315.57463 105.73292 329.66652 329.66652 329.66652

905 906 907 908 909 910 911 912

315.57463 313.73344 320.47588 320.47588 315.57463 313.73344 315.57463 320.47588

913 914 915 916 917 918 919 920

315.57463 271.23604 315.57463 315.57463 113.94087 329.66652 313.73344 329.66652

921 922 923 924 925 926 927 928

320.47588 89.21389 128.00000 329.66652 329.66652 320.47588 128.00000 315.57463

929 930 931 932 933 934 935 936

329.66652 329.66652 329.66652 313.73344 315.57463 121.31354 128.00000 329.66652

937 938 939 940 941 942 943 944

315.57463 75.55022 320.47588 329.66652 315.57463 329.66652 320.47588 315.57463

945 946 947 948 949 950 951 952

329.66652 329.66652 329.66652 315.57463 121.31354 329.66652 320.47588 329.66652

953 954 955 956 957 958 959 960

313.73344 313.73344 329.66652 271.23604 271.23604 121.31354 315.57463 113.94087

961 962 963 964 965 966 967 968

315.57463 260.14155 329.66652 315.57463 271.23604 329.66652 320.47588 113.94087

969 970 971 972 973 974 975 976

320.47588 271.23604 329.66652 121.31354 75.55022 320.47588 320.47588 329.66652

977 978 979 980 981 982 983 984

329.66652 342.38228 121.31354 329.66652 315.57463 329.66652 121.31354 342.38228

985 986 987 988 989 990 991 992

315.57463 271.23604 271.23604 260.14155 121.31354 320.47588 329.66652 121.31354

993 994 995 996 997 998 999 1000

329.66652 271.23604 121.31354 329.66652 271.23604 89.21389 271.23604 329.66652